Proceedings of the workshop on challenges in the management of large corpora and big data and natural language processing (CMLC-5+BigNLP) 2017 including the papers from the web-as-corpus (WAC-XI) guest section. Birmingham, 24 july 2017

Edited by: Bański, Piotr ; Kupietz, Marc ; Lüngen, Harald ; Rayson, Paul ; Biber, Hanno ; Breiteneder, Evelyn ; Clematide, Simon ; Mariani, John ; Stevenson, Mark ; Sick, Theresa

Piotr Bański, Marc Kupietz, Harald Lüngen, Paul Rayson, Hanno Biber, Evelyn Breiteneder, Simon Clematide, John Mariani, Mark Stevenson, Theresa Sick (editors)

Proceedings of the Workshop on *Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017*

including the papers from the

Web-as-Corpus (WAC-XI)

guest section

Birmingham, 24 July 2017

Challenges in the Management of Large Corpora and Big Data and Natural Language Processing 2017

Workshop Programme 24 July 2017

WAC-XI guest session (11.00 - 12.30)

Convenors: Adrien Barbaresi (ICLTT Vienna), Felix Bildhauer (IDS Mannheim), Roland Schäfer (FU Berlin)

Chair: Stefan Evert (Friedrich-Alexander-Universität Erlangen-Nürnberg)

Edyta Jurkiewicz-Rohrbacher, Zrinka Kolaković, Björn Hansen
Web Corpora – the best possible solution for tracking rare phenomena in underresourced languages – Clitics in Bosnian, Croatian and Serbian

Vladimir Benko – *Are Web Corpora Inferior? The Case of Czech and Slovak*

Vit Suchomel – *Removing Spam from Web Corpora Through Supervised Learning Using FastText*

Lunch (12:30-13:30)

CMLC-5+BigNLP main section: (13:30 –17:00)

Welcome and Introduction 13:30-13:40

National Corpora Talks 13:40 – 15:40

Dawn Knight, Tess Fitzpatrick, Steve Morris, Jeremy Evas, Paul Rayson, Irena Spasic, Mark Stonelake, Enlli Môn Thomas, Steven Neale, Jennifer Needs, Scott Piao, Mair Rees, Gareth Watkins, Laurence Anthony, Thomas Michael Cobb, Margaret Deuchar, Kevin Donnelly, Michael McCarthy, Kevin Scannell,
Creating CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes – The National Corpus of Contemporary Welsh)

David McClure, Mark Algee-Hewitt, Douris Steele, Erik Fredner and Hannah Walse,
Organizing corpora at the Stanford Literary Lab

Marc Kupietz, Andreas Witt, Piotr Bański, DanTufiş, DanCristea, TamásVáradi,
EuReCo - Joining Forces for a European Reference Corpus as a sustainable base for cross-linguistic research

John Kirk, Anna Čermáková,
From ICE to ICC: The International Comparable Corpus

Coffee break 15.00-15.20

Harald Lüngen, Marc Kupietz,
CMC Corpora in DeReKo

Technology Talks 15:40-16:40

John Vidler, Stephen Wattam,
Keeping Properties with the Data: CL-MetaHeaders – An Open Specification

Andreas Dittrich,
Intra-connecting a small exemplary literary corpus with semantic web technologies for exploratory literary studies

Radoslav Rábara, Pavel Rychlý ,Ondřej Herman,
Accelerating Corpus Search Using Multiple Cores

Wrap-up discussion 16:40-17:00

CMLC-5+BigNLP Organisers and Editors

Piotr Bański, Marc Kupietz, Harald Lüngen	Institut für Deutsche Sprache, Mannheim
Hanno Biber, Evelyn Breiteneder	Institute for Corpus Linguistics and Text Technology, Vienna
Simon Clematide	Institute of Computational Linguistics, Zurich
John Mariani, Paul Rayson	Lancaster University, UK
Mark Stevenson	Sheffield University, UK

CMLC-5+BigNLP Programme Committee

Laurence Anthony	Waseda University, Japan
Alistair Baron	Lancaster University, UK
Felix Bildhauer	IDS Mannheim
Damir Ćavar	Indiana University, Bloomington
Matt Coole	Lancaster Univeristy, UK
Dan Cristea	Romanian Academy, Institute for Computer Science-Iaşi, "Alexandru Ioan Cuza" University of Iaşi
Tomaž Erjavec	Jožef Stefan Institute
Alexander Geyken	Berlin-Brandenburgische Akademie der Wissenschaften
Johannes Graën	University of Zurich
Andrew Hardie	Lancaster University
Serge Heiden	ENS de Lyon
Miloš Jakubíček	Lexical Computing Ltd.
Dawn Knight	Cardiff University, UK
Michal Křen	Charles University, Prague
Sandra Kübler	Indiana University, Bloomington
Jochen Leidner	Thomson Reuters, UK
Rao Muhammad Adeel Nawab	COMSATS, Pakistan
Piotr Pęzik	University of Łódź
Laura Irina Rusu	IBM Australia
Roland Schäfer	FU Berlin
Roman Schneider	IDS Mannheim
Gandhi Sivakumar	IBM Australia
Irena Spasić	Cardiff University, UK
Marko Tadić	University of Zagreb, Faculty of Humanities and Social Sciences
Dan Tufiş	Institute for Artificial Intelligence Mihai Drăgănescu, Bucharest
Tamás Váradi	Research Institute for Linguistics, Hungarian Academy of Sciences
Andreas Witt	University of Cologne, IDS Mannheim, University of Heidelberg
Amir Zeldes	Georgetown University, USA

CMLC-5+BigNLP Homepage: http://corpora.ids-mannheim.de/cmlc-2017.html

Table of contents

Author Index

Preface

The CMLC+BigNLP workshop is a joint initiative of two teams who have decided to join forces for the purpose of organizing an event co-located with Corpus Linguistics 2017 in Birmingham. The meeting continues the successful series of "Challenges in the Management of Large Corpora" events (previously hosted at CL and LREC conferences) and is at the same time the second event in the Big-NLP series, inaugurated last year at the IEEE Big Data 2016 conference. This year, we wish to explore together common areas of interest across a range of issues in language resource management, corpus linguistics, natural language processing and data science.

An increasing amount of text is available in digital format: more historical archives are being digitised, more publishing houses are opening their textual assets for text mining, and many billions of words can be quickly sourced from the web and online social media. The resulting large textual datasets are used across a number of disciplines to answer a wide range of research questions. In order for these datasets to be maximally useful, careful consideration needs to be made regarding their design, collection, cleaning, encoding, annotation, storage, retrieval and curation.

A number of key themes and questions emerge of interest to the contributing research communities: (a) is having more data always better? (b) is the full range of text types available online and what quality issues should we be aware of? (c) what infrastructures and frameworks are being developed for the efficient storage, annotation, analysis and retrieval of large datasets? (d) what affordances do visualisation techniques offer for the exploratory analysis approaches of corpora? (e) what are the key legal and ethical issues related to the use of large corpora? The present volume contains reports on the current stage of several national large-corpus initiatives and reflects the current thinking on the issues of management and exploitation of large datasets.

Intra-connecting an exemplary literary corpus with semantic web technologies for exploratory literary studies

Andreas Dittrich
Academiae Corpora (Austrian Academy of Sciences) / Sonnenfelsgasse 19/8, 1010 Vienna
andreas.dittrich@oeaw.ac.at

Abstract

Many (modernist) works of literature can be understood by their associativeness, be it constructed or "free". This network-like character of (modernist) literature has often been addressed by terms like "free association", connotation", "context" or "intertext". This paper proposes an experimental and exemplary approach to intra-connect a literary corpus of the Austrian writer Ilse Aichinger with semantic web-technologies to enable interactive explorations of word-associations.

1 Introduction

> "Nearly all poetry is strongly associative." (Cuddon, 2013, p. 58)

Large corpora are rich corpora. Following the etymological routes of the word, Latin *largus* does not mean "thick" and "coarse", like the root of the word "great", but "plentiful" and "abundant". The difference between large and small corpora thus is not the simple measure of quantitative size, but the question of how to deal with it: a methodological question.

For John Sinclair, for whom "the difference [between small and large corpora] must be methodological" (Sinclair, 2001, p. xi), "[t]he main virtue of being large in a corpus is that the underlying regularities have a better chance of showing through the superficial variations" (Sinclair, 2004, 189). In the field of literary studies this "underlying regularities" can be various: a theme, plot, motif, sujet and fabula, device, meaning, rhetoric, trope, style, metric, sound or others. But all these refer to a specific text, which can be gathered as a corpus — and, as a digital corpus, analysed with computational methods. A traditional approach of analysing texts, called "close reading", has been extended by a method roughly labelled as "distant reading", which tries to analyse not just one text, but a plenty. If one doesn't understand these terms as opposites, but as different moments of the same process, one can get to read texts close via distant readings and vice versa (Jänicke et al., 2015; Scrivner and Davis, 2017; Jockers, 2013), more or less as Hans-Georg Gadamer describes the structure of understanding as a "circle of whole and part" (Gadamer, 2004, p. 302–5) (although "whole" probably is a hole).

This constant moving between macro- and micro-structure, requires an interactive workframe without delay, which, depending on the size of the corpus, can be difficult to obtain and the idea of lessen the corpus may occur. One of the apparently most natural processes before or after Natural Language Processing (*NLP*) is the exclusion of stop-words. This crucial intervention alters the corpus drastically and deletes merely seemingly 'meaningless' words like the copula "and", which could be a decisive stylistic factor for an author. Such filtering methods, which are important for making corpora suitable for analysis, reduce the richness and thereby the largeness of a corpus. Usual literary corpora may not reach the quantitative size of comparable corpora from Linguistics in their quantitative scale, but may tend there, when they focus on connections between words.

In the following, I want to discuss a project that deals with texts of a specific author (Ilse Aichinger), whose corpus, which we finished to build in TEI-XML[1] (Text Encoding Initiative, Extensible Markup Language), is small in quantitative size (about

[1] We is a group of students under supervision of Christine Ivanovic from the Institute of Comparative Literature at the University of Vienna and Hanno Biber from Academiae Corpora at the Austrian Academy of Sciences: Marlene Csillag, Katharina Godler, Mathias Müller, Katrin Rohrbacher, Gilbert Waltl and myself.

400.000 tokens), but rich regarding its literary interconnectivity (Fässler, 2011; Thums, 2013; Pelz, 2009; Markus, 2015). After discussing the work of Ilse Aichinger and which problems occurred to us in the process of annotating place-names, I want to propose an interactive visualisation-method, which is based on technologies of the semantic web. For this purpose the XML-files had to be converted to a RDF-format (`Resource Description Framework`). Finally, I present an exemplary, very short study of three words from the corpus in an open-source visualization-framework, named "RelFinder" (Heim et al., 2010). This not only offers 'new' questions for the field of literary studies, but enables us to see other connections between texts, discovering mediating terms and second-order mediations.

2 Places in the corpus : `aichinger`

Places play an eminent role in the writing of Ilse Aichinger (1921–2016). In order to protect her mother, whom the Nazi-regime labelled as "half-Jewish", she did not emigrate from National Socialist Vienna, where she survived second world war. Places trigger a process of remembrance, and thus "the vanished" acquire a literary presence in their absence (Fässler, 2011, p. 26). The places 'touch' Aichinger in her present being (Thums, 2013, p. 193): "The places, which we looked at, look at us" (Aichinger, 2001, my translation, AD), as she writes in a short text. But place-names are not simply uttered or just staging the scene, they also "carry the plot", as she once noted herself (Aichinger, 1991b, p. 179, my translation, AD).

The difficulties in the annotating-process have been diverse and can be summed up in the question: How to define a literary place? This question arose probably because of the very different 'styles' Aichinger exhibits in her entire oeuvre, which spans over 60, transformative years, from her first published text 1945 to her last one 2005. The annotating group faced texts, where very different place-types turned up: fictitious place-names, moving places, acting places, existing place-names, which do not refer to their real place-reference, but also place-names that can be located on a traditional map. The group agreed, that, at least as a first step, only place-names should be annotated, which can be located on a map. Additionally to a light TEI-encoding (with page-breaks, line-breaks, divisions and headings with corresponding publication dates and genre, paragraphs, stage directions, speaker and speeches, line-groups and lines), place-names were manually annotated by using the "referencing string"-tag (`rs`) with the attribute (`type`) "place" (see Fig. 1).

```
<body>
<div n="1" ana="ed19480000 eb19480000" type="prose">
<head>Die größere Hoffnung</head>
<div type="chapter">
<pb facs="../files/Aichinger-Hoffnung_n0009.tif" n="9"/>
<head>Die große Hoffnung</head>
<p>Rund um das <rs type="place">Kap der Guten Hoffnung</rs> wurde das Meer<lb/>dunkel.
Die Schiffahrtslinien leuchteten noch einmal auf und<lb/>erloschen. Die Fluglinien
sanken wie eine Vermessenheit.<lb/>Ängstlich sammelten sich die Inselgruppen. Das
Meer<lb/>überflutete alle Längen- und Breitengrade. Es verlachte das<lb/>Wissen der
Welt, schmiegte sich wie schwere Seide gegen das<lb/>helle Land und ließ die <rs
type="place">Südspitze von <rs type="place">Afrika</rs></rs> nur wie eine<lb/>Ahnung
im Dämmern. Es nahm den Küstenlinien die<lb/>Begründung und milderte ihre
Zerrissenheit.</p><p>Die Dunkelheit landete und bewegte sich langsam gegen<lb/>Norden.
Wie eine große Karawane zog sie die Wüste hinauf,<lb/>breit und unaufhaltsam. Ellen
schob die Matrosenmütze aus<lb/>dem Gesicht und zog die Stirne hoch. Plötzlich legte
sie die<lb/>Hand auf das <rs type="place">Mittelmeer</rs>, eine heiße kleine Hand.
Aber es half<lb/>nichts mehr. Die Dunkelheit war in die <rs type="place">Häfen von <rs
type="place">Europa</rs></rs><lb/>eingelaufen.</p><p>Schwere Schatten sanken durch die
weißen Fensterrahmen.<lb/>Im Hof rauschte ein Brunnen. Irgendwo verebbte ein
Lachen.<lb/>Eine Fliege kroch von <rs type="place">Dover</rs> nach <rs
type="place">Calais</rs>.</p>
```

Figure 1: Exemplary screenshot of a TEI-XML-file.

This resulted in about 1.800 references to real places. Previous scholarly works have not seen this multitude of references in the text (Schmid-Bortenschlager, 2001). Moreover the text with the most quantity and diverse real-place-references ("Nachricht vom Tag") is, surprisingly or not, one which among Aichinger-scholars is very rarely discussed (see figure 2). Further it could be shown, that real-place-references are not exclusive but predominant in Aichinger's later work (see Fig. 3). Previously similar results have been shown with simple text-query-statistics (Frank and Dittrich, 2015, p. 52–53).

Although promising techniques of automatic place-name-recognition are in development (Bornet and Kaplan, 2017) the annotations have been made manually. The special challenge in this case was, to get to terms relating to the different types of place-references. Mahler and Dünne proposed to differ between "topography" and "topology" (Dünne and Mahler, 2015, p. 6). Topographical entities operate in a semantic reference system and can therefore be mapped. Topological entities operate in a syntactic relation system and are therefore able to get located in a network. It is not easy to differ between those two categories in every case. The notion of "Dover" for example can refer both to the real place in the south of Great Britain and be an empty signifier not referring to anything at all (Aichinger, 1991a). Only out of this undecidable entanglement the playful meaning of the text arises. To grasp the interwoven conjunc-

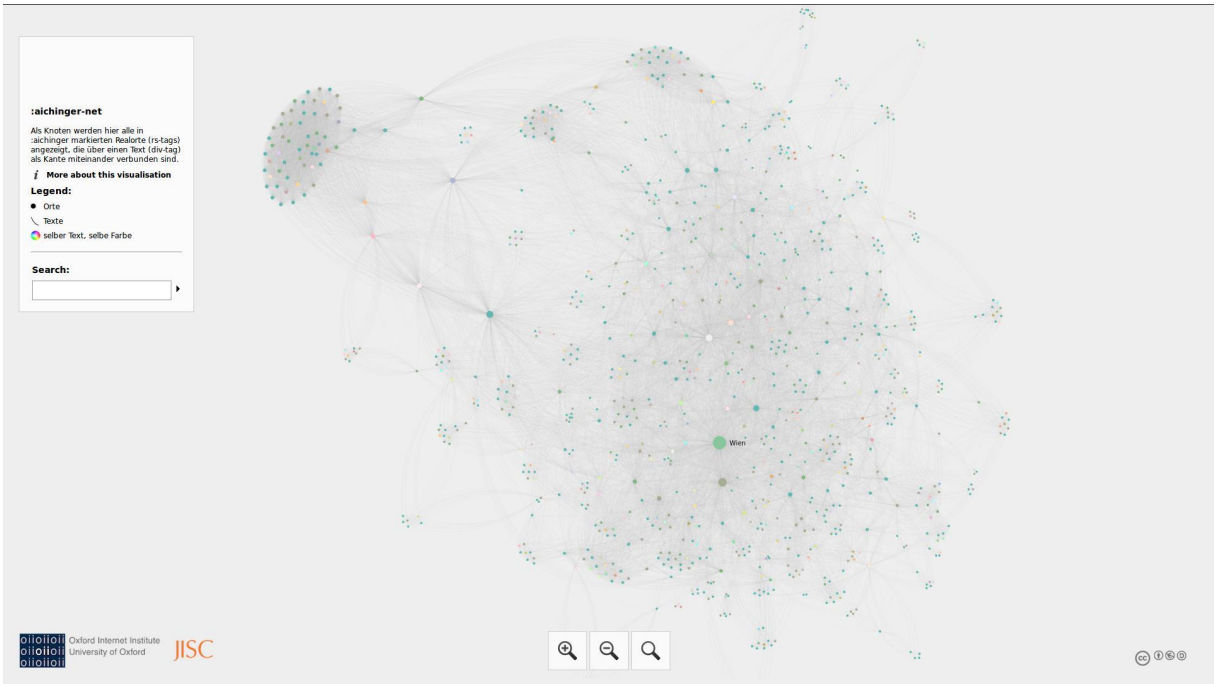

Figure 2: Network-view of all tagged place-names: nodes are place-names, edges are text-divisions. The cluster in the upper left corner represents the place-names in the text "Nachricht vom Tag". This graph can be explored online: `http://homepage.univie.ac.at/andreas.dittrich/aichinger-rsnet`. Visualization made with *Gephi* and *sigmajs.org*.

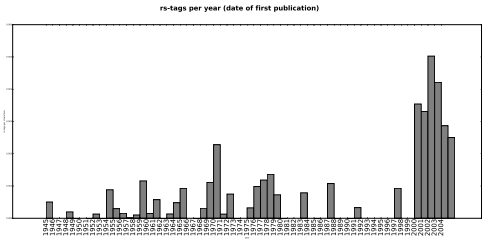

Figure 3: A time-based view (first publication) of place-name-frequency per characters.

tion of topographic and topologic entities it is imperative to first address them separately. But although this distinction is useful for first steps, it does not exhaust the many possibilities of place-references in literature. To name just a few, which we encountered:

- place-names referring to real places and which are mappable;
- common place-names like "kitchen" or "park";
- fictional places like a "fan";
- and place-names that simply cannot be located like "Port Sing".

3 Towards an exploratory framework

Heinz Schafroth suggested to read the texts of Aichinger "associative" (Schafroth, 1976, p. 130), that is to say: reading the intra-connectivity of the different texts. Following this proposal, we can represent the texts as a network and make it explorable as such. Simple methods in Corpus-Studies work with types or word-forms and answer questions like: where, how often and in which context can I find a specific word in the corpus. Even queries about co-occurrences of words are possible. But how about words that share the same co-occurrences, but not the same words?

Say, for example, the word "Vater" occurs in a set of texts A, "Mutter" in a set of texts B. Let us call the overlapping of shared words AB. Now, there are texts, which share words with the set of text A, not B, but share words with a set of texts C, which share words with B (see Fig. 4). This set of texts C can be interesting for analysis — and maybe this is, what Peirce called "abductive reasoning" —, but it would be difficult to reach within the boundaries of conventional queries.

A SPARQL-server (Apache Jena Fuseki), which stores RDF-files, is used. If a simple RDF-turtle-file would contain the following informa-

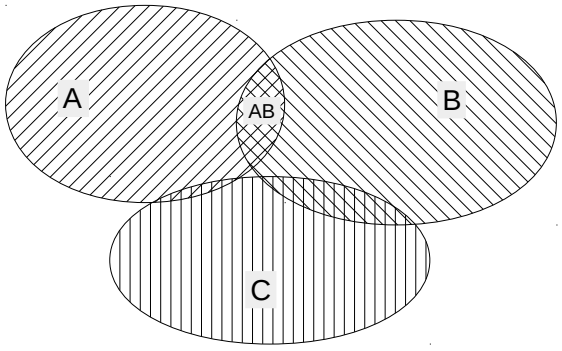

Figure 4: Illustration of set of texts.

tion:

```
@prefix word: <ia.net/words#>.
@prefix rel:  <ia.net/relat#>.
word:father rel:str  "Father".
word:father rel:tr word:vater.
word:vater  rel:str   "Vater".
```

a simple query could look like this:

```
SELECT ?o1 ?p2 ?o2
WHERE {
  ?m1 ?p1 ?o1 .
  ?m1 ?p2 ?m2 .
  ?m2 ?p3 ?o  .
}
```

which would result in a formatted output like this:

```
"Father" rel:tr "Vater"
```

To this end, the TEI-XML-files have to be converted in RDF, for which an special Python-program had to be written. In the RDF-file all words, which are in the same division of text, are connected with each other (this is useful, because Aichinger mainly wrote short texts); date-, genre- and place-annotations are linked to the division (see Fig. 5).

To explore such a network not only by its "most frequent" or "most linked" terms one needs to be able to move inside of this network intuitively. Text-corpora of about 400.000 tokens could result in a network of about 81 billion connections. But immediate interactive and visual exploration of these networks is needed. It should be possible to alter the graph (for example to add, drag or remove certain nodes or edges) and see the results without delay. Developed for the so-called semantic web,

```
iawork:1_Die_große_Hoffnung rdfs:label "DIE GROSSE HOFFNUNG";
   lit:date "1948";
   lit:genre "prose";
   lit:inbook "Werke. Die größere Hoffnung";
   lit:pages "S. 9 ff.";
   lit:abstract "TITEL: Die große Hoffnung,\nERSTPUBLIKATION: 1948,\nGENRE: prose,
\nIN: Werke. Die größere Hoffnung, S. 9 ff.";
   lit:has iaplace:tipperary, iaplace:dover, iaplace:calais,
iaplace:hafen_von_europa, iaplace:rund_um_die_kap_die_gut_hoffnung,
iaplace:südspitze_von_afrika, iaplace:europa, iaplace:hamburg, iaplace:mittelmeer,
iaplace:tschechisch, iaplace:freiheitsstatue, iaplace:hawaii,
iaplace:kap_die_gut_hoffnung, iaplace:amerika, iaplace:pazifisch_ozean,
iaplace:afrika, iaplace:sibirien;
   lit:has iaword:ans, iaword:regnen, iaword:können, iaword:machen, iaword:mund,
iaword:runden, iaword:tanzen, iaword:hereinkommen, iaword:aufrichten, iaword:wild,
iaword:ecke, iaword:hemd, iaword:schrei, iaword:herab, iaword:zerren,
iaword:rettungsgürtel, iaword:bein, iaword:eiskalt, iaword:grenzenlos,
iaword:finden, iaword:ach, iaword:schatten, iaword:tod, iaword:ausgestreckt,
iaword:sprechen, iaword:schiffahrtslinie, iaword:blass, iaword:kirchenstufen,
iaword:mittelmeer, iaword:schreiben, iaword:zweit, iaword:wunderbar, iaword:seit,
```

Figure 5: Exemplary screenshot of a RDF-file.

which works with structured data, the open-source software "RelFinder" offers a suitable framework to make such interactive queries (see `http://relfinder.visualdataweb.org`).

See Fig. 6 for an example of how the words "vater", "mutter" and "kind" are related: five texts appear in the center of the graph, which share the three words. What may catch the eye of an Aichinger-scholar is the centrality of the term "augenblick" (blink of an eye, moment), which is central to her concept of "hope" (Thums, 2013, p. 193–196), which leads to her novel "The Greater Hope" (Aichinger, 2016). And it not only seems to connect all other nodes but it connects most of the nodes, which are connected with the three searched ones. "Die Zumutungen des Atmens" for example is connected to "mutter", "vater" and "kind", but also to "augenblick", which it shares with "Die Spiegelgeschichte". (The same can be said for "Die größere Hoffnung" (chapter), "Eliza Eliza" (text) and "Die Schwestern Jouet" (drama). The only Text, which does not share "augenblick", but all other words, is "Bin noch immer positiv!".) Although some text do not share all the searched words, many of them share the word "augenblick". It is possible to lessen the graph's output by different mechanisms. See Fig. 7 where all relations, that are direct only, are faded out. The nodes "augenblick" and "Die Spiegelgeschichte" stay in the center and suggest a high connectivity.

4 Conclusion and Discussion

Although the presented approach is in developement and not all possibilities are exhausted yet, it could be shown what the basic idea enables: Finding maybe unexpected connections between texts and by this, enabling new insights into already known connections and discovery of unknown interrelations. The concept of "Augenblick" has already been in the focus of Aichinger-scholars, but

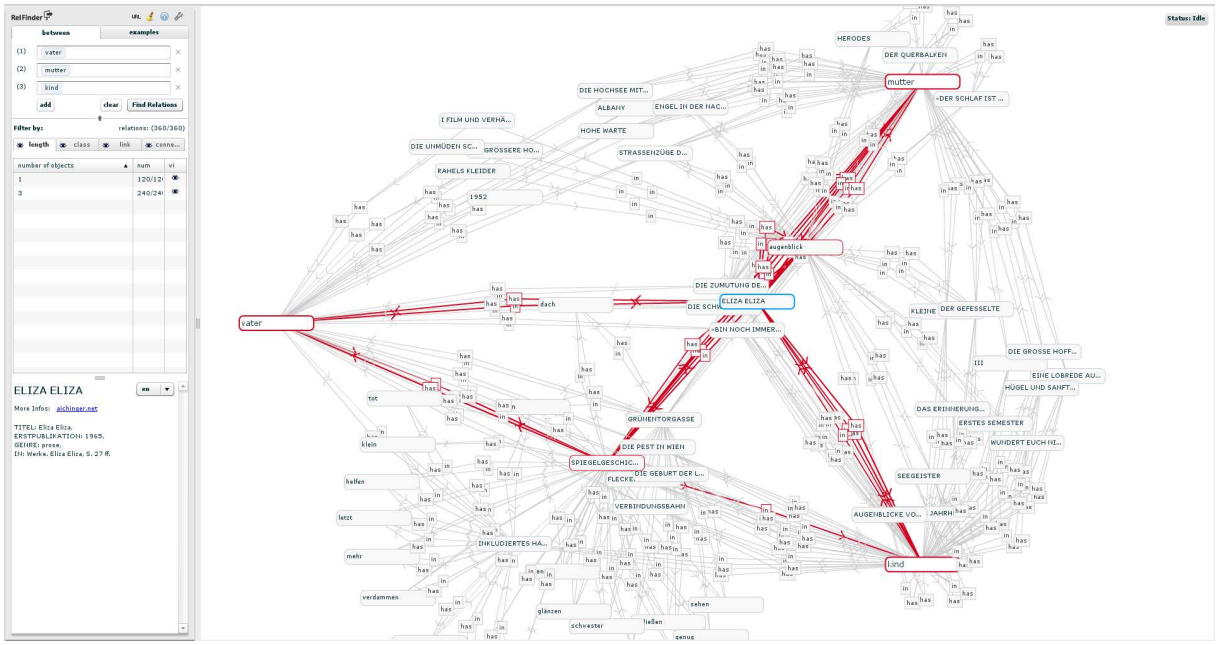

Figure 6: A view of all associations of “vater”, “mutter” and “kind” in the corpus `:aichinger` with *RelFinder*.

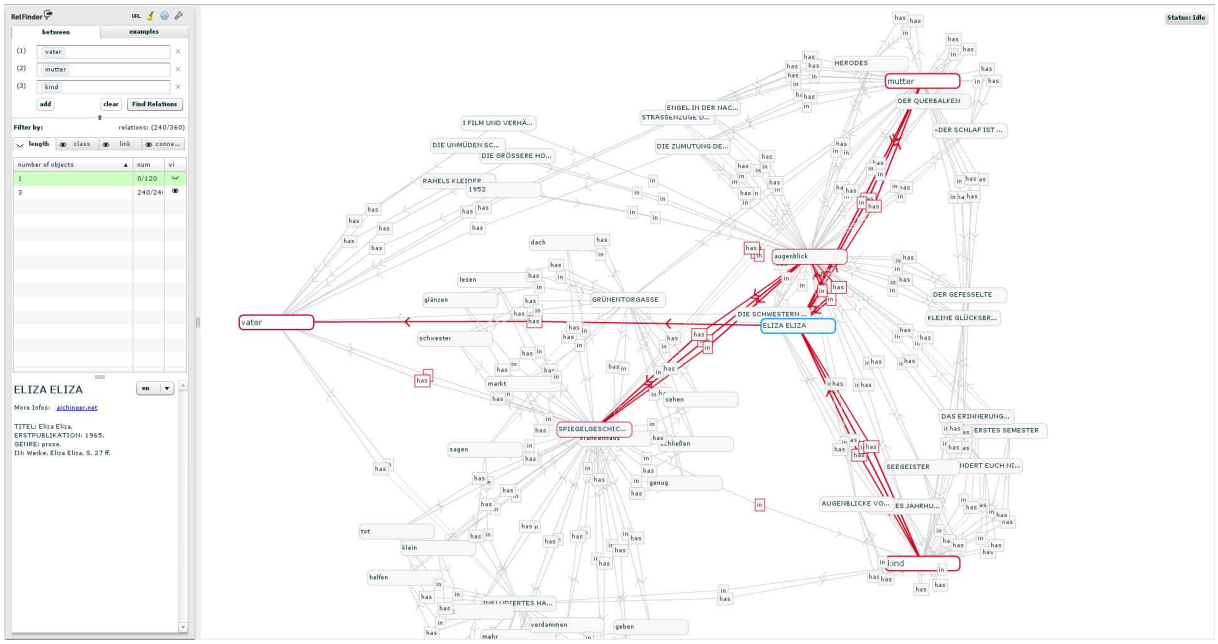

Figure 7: A selected view of the associations: Only those nodes are displayed whose connections are not only direct.

not in the perspective of its relatedness to other texts and words.

A crucial point in this types of visualisations is the eclipse of the dimension of time. The graph seems to suggest, that these words are used in a timeless space. I tried to adjust this by listing meta-data with first publications-date and genre on the left side of the screen. To make it easier to read the texts in their entire context, the book and page-number, where the texts can be found, are listed.

Of course some methodological problems do persist in this approach. One, that troubles me basically, is that this approach seems to assume that words mean the same in different contexts. But they don't. Not even in, or maybe most notably not in literature. By unifying different singular occurrences of a word to one word-type, the singular use in a singular context gets covered. One has to be vigilant to not level important differences. Ilse Aichinger wrote a text called "Hemlin", which performs the variability of words by using the untranslatable (or exactly translatable) word "Hemlin" in various ways (Markus, 2015, p. 89–90) and questions the – sometimes undue – unifying drive of scientific methods: "Hemlin must be a monument, round, makes trouble." (Wolf and Hawkey, 2010, p. 191). Hemlin.

References

Ilse Aichinger. 1991a. *Schlechte Wörter*, volume 4. Fischer.

Ilse Aichinger. 1991b. *Zu keiner Stunde. Szenen und Dialoge*, volume 7. Fischer.

Ilse Aichinger. 2001. *Kurzschlüsse*. Edition Korrespondenzen.

Ilse Aichinger. 2016. *The Greater Hope*. Königshausen & Neumann.

Cyril Bornet and Frédéric Kaplan. 2017. A simple set of rules for characters and place recognition in french novels. *Frontiers in Digital Humanities*, 4:6.

John Anthony Cuddon. 2013. *A Dictionary of Literary Terms and Literary Theory*. Wiley-Blackwell, 5 edition.

Jörg Dünne and Andreas Mahler. 2015. Einleitung. In Jörg Dünne and Andreas Mahler, editors, *Handbuch Literatur & Raum*, pages 1–11. Walter de Gruyter.

Andrew U. Frank and Andreas Dittrich. 2015. Flexible annotation of digital literary text corpus with rdf. In Francesco Mambrini, Marco Passarotti, and Caroline Sporleder, editors, *Proceedings of the Workshop on Corpus-Based Research in the Humanitie*, pages 49–58. Institute of Computer Science. Polish Academy of Sciences.

Simone Fässler. 2011. *Von Wien her, auf Wien hin*. Böhlau.

Hans-Georg Gadamer. 2004. *Truth and Method*. Bloomsbury, 2 edition.

Philipp Heim, Steffen Lohmann, and Timo Stegemann. 2010. Interactive relationship discovery via the semantic web. In *Proceedings of the 7th Extended Semantic Web Conference (ESWC 2010)*, volume 6088 of *LNCS*, pages 303–317. Springer.

Matthew L. Jockers. 2013. *Macroanalysis*. Topics in the Digital Humanities. University of Illinois Press.

Stefan Jänicke, Greta Franzini, Muhammad Faisal Cheema, and Gerik Scheuermann. 2015. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In R. Borgo, F. Ganovelli, and I. Viola, editors, *Eurographics Conference on Visualization (EuroVis) - STARs*. The Eurographics Association.

Hannah Markus. 2015. *Ilse Aichingers Lyrik. Das gedruckte Werk und die Handschriften*. De Gruyter.

Annegret Pelz. 2009. Spracharbeit in meeresnähe. In Ingeborg Rabenstein-Michel, François Rétif, and Erika Tunner, editors, *Ilse Aichinger – Misstrauen als Engagement?*, pages 63–72. Könighausen & Neumann.

Heinz F. Schafroth, 1976. *Die Dimensionen der Atemlosigkeit*, pages 129–133. Fischer.

Sigrid Schmid-Bortenschlager. 2001. Die topographie ilse aichingers. In *Was wir einsetzen können, ist Nüchternheit*, pages 179–188. Königshausen & Neumann.

Olga Scrivner and Jefferson Davis. 2017. Interactive text mining suite: Data visualization for literary studies. In Thierry Declerck and Sandra Kübler, editors, *Proceedings of the Workshop on Corpora in the Digital Humanities (CDH 2017)*, pages 29–38.

John Sinclair. 2001. Preface. In Mohsen Ghadessy, Alex Henry, and Robert L. Roseberry, editors, *Small Corpus Studies*, page vii–xvi. Benjamin Publishing.

John Sinclair. 2004. *Trust the Text: Language, Corpus and Discourse*. Routledge.

Barbara Thums. 2013. Zumutungen, ent-ortungen, grenzen. In Doerte Bischoff and Susanne Komfort-Hein, editors, *Literatur und Exil*, pages 183–209. De Gruyter.

Uljana Wolf and Christian Hawkey. 2010. Notizen beim Übersetzen von aichingers hemlin. *Berliner Hefte zur Geschichte des literarischen Lebens*, (9):188–191.

From ICE to ICC:
The new *International Comparable Corpus*

John Kirk
4 Parkvue Manor
Belfast, BT5 7TJ
Northern Ireland
jk@etinu.com

Anna Čermáková
Charles University Prague
Institute of the Czech National Corpus
nám. J. Palacha 2
116 38, Praha 1
anna.cermakova@ff.cuni.cz

Abstract

This paper outlines the broad research context and rationale for a new international comparable corpus (ICC). The ICC is to be largely modelled on the text categories and their quantities the *International Corpus of English* with only a few changes. The corpus will initially begin with nine European languages but others may join in due course. The paper reports on those and other agreements made at the inaugural planning meeting in Prague on 22-23 June 2017. It also sets out the project's goals for its first two years.

1 International Corpus of English (ICE) project

There is broad agreement that the *International Corpus of English* (ICE) project has been highly successful because it has facilitated numerous systematic comparisons of L1 and L2 national varieties of English worldwide. Those comparisons encompass the lexical and morpho-syntactic structural levels, as well as comparisons of discourse types and written registers (cf. e.g. Greenbaum, 1996; Hundt and Gut, 2012; Aarts et al., 2013; and the papers in the Special Issues of *World Englishes* vol. 15(1) (1996) and vol. 36(3) (2017), to mention but a few key studies).

ICE does not sample populations, nor does it relate national component sizes proportionately to the size of the population. Rather ICE is entirely text-based, being organized around text categories and the quantity which has been designated for each one, with each corpus following the same pattern regardless of population size. Each corpus thus amounts to its text collection, no matter whether it is the USA, with a population of 321.4 million, or Malta, with its population of 419,000. It's the identical nature and quantities of that collection which allow for the comparability of the component corpora. No small part of the success of ICE rests with the fact that for each national variety there has been chosen a set of spoken and written text categories which are deemed to be representative of each national variety: 15 discourse situations (totalling 60%) (see Table 1) and 17 written registers (totalling 40%) (see Table 2). The importance of retaining those categories for any second generation ICE corpus for comparability was confirmed in a major review of the ICE project in May 2017.

SPOKEN TEXTS			
DIALOGUE (180)	**Private (100)**	direct conversations	90
		distanced (telephone) conversations	10
	Public (80)	class lessons or seminars	20
		broadcast discussions	20
		broadcast interviews	10
		parliamentary debates	10
		legal cross-examinations	10
		business transactions	10
MONOLOGUE (120)	**Unscripted (70)**	spontaneous commentaries	20
		unscripted speeches	30
		demonstrations	10

		legal presentations	10
	Scripted (50)	broadcast news	20
		broadcast talks	20
		speeches (not broadcast)	10

Table 1: Spoken text categories in ICE.

The spoken texts are categorized by a principled, top-down approach with regard to the speech situation: whether there is one speaker or more than one; whether the speech is public or private; and whether the speech is scripted or spontaneous. The final choice is based largely on functional domain, such as broadcasting, parliament, education, or the law courts. As Table 1 shows, most categories are collected in similar quantities, except private, face-to-face conversation, which predominate, not least because they are regarded as the quintessential form of spoken interaction. However, public speech accounts for two-thirds of all spoken texts (200/300 texts).

In these ways, although they are not without criticism, ICE has come to represent a fair sampling of all the major spoken and written varieties of English in the present day and throughout the world, particularly in L1 countries.

WRITTEN TEXTS			
NON-PRINTED (50)	**Non-professional writing (20)**	student untimed essays	10
		student examination	10
	Correspondences (30)	social letters	15
		business letters	15
PRINTED (150)	**Informational (learned) (40)**	humanities	10
		social sciences	10
		natural sciences	10
		technology	10
	Informational (popular) (40)	humanities	10
		social sciences	10
		natural sciences	10
		technology	10
	Informational (reportage) (20)	press news reports	20
	Instructional (20)	administrative/regulatory prose	10
		skills/hobbies	10
	Persuasive (10)	press editorials	10
	Creative (20)	novels/short stories	20

Table 2: Written text categories in ICE.

The written texts are similarly categorized by a principled, top-down approach with regard to the register situation: whether the text has been printed or not; and what its primary function is. Cutting across two of the main informational functions are domain choices. There are also two types of writing from newspapers: reporting as a further instance of informational writing; and editorials as an instance of persuasive writing. Printed texts account for three-quarters of all written texts (150/200).

2 Contrastive (corpus) linguistics

While ICE has been developing over the last thirty years or so, spoken and/or written corpora have been compiled for other languages (cf. list of non-English corpora in e.g. O'Keeffe et al. (2007, 294-296) or the non-English corpora discussed in Xiao (2008) or Ostler (2008)). Xiao makes comparisons with corpora of English: for instance, the *Polish National Corpus* replicates the structure of the *British National Corpus* (Xiao, 2008, 387), as does, to an extent, the *Czech National Corpus* (Čermák, 1997), which contains spoken texts similar to those of demographically sampled component of BNC (Xiao, 2008, 388-389; Čermák, 2009). However, no corpus of another language appears to be composed with the range and balance of spoken and written text categories and quantities of texts as contained within the ICE corpus. The existing corpora in various languages are generally compiled on very different principles and thus do not allow direct cross-linguistic contrastive comparisons.

Corpus-based contrastive studies are a growing research area and researchers have voiced need for more rigorous analytical framework (e.g. Aijmer et al., 1996; Altenberg and Granger, 2002; Marzo et al., 2012; Aijmer and Altenberg, 2013; Altenberg and Aijmer,

2013; Ebeling and Ebeling, 2013). The majority of contrastive studies are being carried out on two languages only (and very often one of the compared languages is English), one of the reasons being the lack of comparable data. Contrastive analysis relies on two types of data (Granger, 2003): translation (parallel) corpora and comparable corpora (cf. McEnery and Xiao, 2007). While translation corpora contain original (source) texts with their aligned translations, comparable corpora[1] contain original texts in two or more languages that have been selected on comparable criteria for text categories and quantities for each category, such as the *Lancaster Corpus of Mandarin Chinese*, which uses the same sampling frame of the *Lancaster/Oslo-Bergen Corpus*, or the *Aarhus Corpus of Contract Law* (both cited in McEnery and Hardie, 2012: 19; cf. also e.g. Sharoff et al., 2014). Comparable corpora are an essential data source to support contrastive analyses, since the translation corpora are usually limited as far as text types are concerned (e.g. Johansson, 2007; Mauranen, 1999).

3 The International Comparable Corpus (ICC)

3.1 Rationale for ICC

The ultimate goal of this project is the facilitation of contrastive studies between English and other languages involving highly comparable datasets of spoken, written and electronic registers. What we are introducing is not a parallel translation corpus;[2] but rather, it is the creation of an *International Comparable Corpus* (ICC – pronounced to rhyme with *lick*), with as many languages as may wish to come on board. Phase I will start with national, standard(ised) European languages; an expression of interest to collaborate on this project has been expressed for the following languages: Czech, Finnish, French, German, Norwegian, Polish, Slovak, and Swedish. The first collaborative meeting was held on 22–23 June 2017 in Prague[3].

The ICC corpus is based, on the one hand, on the idea that there are plenty of various language data for many languages that could be reused if carefully selected and, on the other, that contrastive analysis very often relies on comparisons with English. Thus the ICC corpus will largely rely on re-using existing language resources and will be modelled for comparability with the ICE family corpora. For the field of contrastive linguistics, a striking and unique feature of each new corpus in ICC will be its substantial spoken component. Such provision of spoken data across 13 or so discourse situations for contrastive analysis among several languages is entirely unique as it will allow the much-needed and unprecedented cross-linguistic corpus-based comparisons of spoken language. Together with balanced data across written registers, ICC will become invaluable for future research[4]. The approach will also allow replicability and comparisons with and between other languages.

3.2 Composition of ICC

Let us now turn to some specifics about the new ICC. Following agreement in Prague, the ICC will broadly follow the composition of the ICE corpus, see Tables 3 and 4, the rationale for those text categories as briefly outlined above being taken largely for granted. Individual texts will comprise approximately 2,000 tokens each, ending with sentences or paragraphs completed (if possible); many texts will be excerpts, derived from a good spread of beginnings, middles and endings of their source texts. If texts are shorter than 2,000 words, composite texts are to be created, to make up the desired total. Within categories, the texts are to be chosen on the basis of the range, spread and diversity of the category or the function which the texts represent. Texts are to post-date 2000, and there are to be no

[1] Terminology may differ, but here we mean by parallel corpora, source language texts aligned to their translations. Comparable corpora may be multilingual as referred to in this article but also monolingual, containing comparable datasets in one language, e.g. non-translated language and translated language such as the *The Translation English Corpus* (TEC) (Baker, 1995), available at http://www.monabaker.com/tsresources/TranslationalEnglishCorpus.htm).

[2] Such as the *English-Swedish Parallel Corpus*, the *English-Norwegian Parallel Corpus* (ENPC), or the *InterCorp* corpus.

[3] We would like to thank the following participants in the ICC planning meeting: Michal Křen (Czech), Oliver Wicher (French), Marc Kupietz (German), Signe Oksefjell Ebeling (Norwegian), Jarle Ebeling (Norwegian), Rafal Gorski (Polish), Radovan Garabík (Slovak), Vladimir Benko (Slovak), and the following for their input and support of the ICC idea: Jarmo Jantunen (Finnish), Dirk Siepmann (French), Christoph Bürgel (French), Sascha Diwersy (French), Thomas Schmidt (German), Mária Šimková (Slovak), and Karin Aijmer (Swedish).

[4] Cf. e.g. studies of English-German contrasts, such as König and Gast (2012), or English-Norwegian contrasts, such as Ebeling and Ebeling (2013).

translations. Ideally, no source text is to be used more than once. Moreover, ICC has decided to drop all non-printed texts (undergraduate essays and letters, totalling 100 texts) and the two spoken categories of legal texts (each 10 texts). However, it has been decided to add a category of (electronic) blogs (50 texts each of 2,000 words) equivalent to the non-printed texts now dropped. The total for the spoken component will be 560,000 words. As far as possible texts are to be selected from existing national resources, to maximise their re-usability and to minimise the effort. ICC is not intended to replicate or compete with national corpora; rather the emphasis is on systematic comparability between and across languages. As with ICE, so will it be for ICC: identical types and quantities of texts will neutralize any population differences between participating countries, whether 81 million for Germany, or 5.2 million for Norway. As ICE is a corpus of English, so ICC is to be a corpus of languages.

	CZE	FIN	FRE	GER	NOR	POL	SLO	SWE
Hum. (acad.)	√		√	√	√	√	√	
Soc. sci. (acad.)	√		√		√	√	√	
Nat. sci. (acad.)	√		√		√	√	√	
Technol. (acad.)	√		√	√	√	√	√	
Hum. (pop.)	√	√	√	√	√	√	√	
Soc. sci. (pop.)	√	√	√	√	√	√	√	
Nat. sci. (pop.)	√	√	√	√	√	√	√	
Technol. (pop.)	√		√	√	√	√	√	
Reportage	√	√	√	√	√	√	√	
Instruct. (admin.)	√	√	√		√	√	√	
Instruct. (hobbies)	√		√		√	√	√	
Press editorials	√			√	√	√	√	
Fiction	√		√	√	√	√	√	
Blogs	√	√	√	√	√	√	√	√

Table 3: Written categories agreed for ICC and their availability from currently identified resources (as of June 2017).

	CZE	FIN	FRE	GER	NOR	POL	SLO	SWE
Direct convers.	√	√	√	√		√	√	√
Telephone convers.			√	√				√
Class (uni) lessons								
Broadcast discussions	√	√	√		√	√	√	
Broadcast interviews	√		√		√	√	√	
Parliament debates			√		√		√	
Business transact.			√	√				√
Spontan. comment.			√					
Unscripted speeches			√				√	
Demonst. (broadc.)								
Broadcast News		√	√		√	√		
Broadcast Talks		√			√	√		
Speeches (not broadcast)			√					

Table 4: Spoken categories agreed for ICC and their availability from currently identified resources (as of June 2017).

Both written and spoken texts are to be marked up in a format conforming to TEI P5 XML,[5] keeping the original characters (multiple dashes, apostrophes etc.). Each text is to be accompanied by metadata in an accompanying header. As, for the spoken component, sound alignment is strongly desired wherever possible, a multi-layer environment will be needed, such as ELAN, in which one-layer will contain the orthographic transcription. Transcription details are to be language-dependent. However, overlaps and pauses are to be included and marked according to TEI. A minimum markup scheme is being drawn up.

Texts are to be annotated with regard to the part of speech (POS) status with POS taggers representing state-of-the-art for each language. As, among the languages, considerable morphological variation exists, another simultaneous tagging layer was considered for mapping language-specific POS annotation schemes onto higher level "universal" schemes

[5] http://www.tei-c.org/Guidelines/P5/

(e.g. 'universal dependencies' or simplified tagset used for the *Aranea* corpora series)[6] to support cross-linguistic comparisons. A further aspiration for the future is for cross-linguistic syntactic tagging and parsing.

The ICC is to be made available through a common search interface with distributed indexes (KorAP).[7] However, there is a preference for ICC components to be downloadable, at least partially[8], and with non-destructive annotation, but that will depend on copyright permissions being cleared in the first instance. As a plan of action, it was decided to re-negotiate licensing of written texts (CC BY-NC), and to choose and attempt to transform the spoken texts into TEI P5 XML format by the end of the first year. By the end of two years, missing spoken texts are to be collected and the pilot written corpus should have been completed.

These, then, are the parameters in terms of which the ICC is to come into being. We are pleased to introduce this exciting, new international corpus. The project welcomes further participation.

References

Aarts B., Close J., Leech G. and Wallis S. (Eds.). 2013. *The Verb Phrase in English*. Cambridge University Press, Cambridge.

Aijmer K. and Altenberg B. (Eds.). 2013. *Advances in Corpus-based Contrastive Linguistics*. John Benjamins, Amsterdam.

Aijmer K., Altenberg B. and Johansson M. (Eds.). 1996. *Languages in Contrast. Papers from a Symposium on Text-based Cross-Linguistic Studies*, Lund, 4–5 March 1994. Lund University Press, Lund.

Aijmer K. and Vandenbergen A.-M. (Eds.). 2006. *Pragmatic Markers in Contrast*. Elsevier, Amsterdam.

Altenberg B. and Aijmer K. (Eds.). 2013. Text-based Contrastive Linguistics. Special Issue of *Languages in Contrast* 13(2).

Altenberg B. and Granger, S. (Eds.). 2002. *Lexis in Contrast: Corpus-based Approaches.* John Benjamins, Amsterdam.

Baker M. 1995. Corpora in Translation Studies. An Overview and Suggestions for Future Research. *Target*, 7(2): 223-243.

Benko V. 2014a. Aranea: Yet Another Family of (Comparable) Web Corpora. In P. Sojka, A. Horák, I. Kopeček and K. Pala (Eds.), *Text, Speech and Dialogue. 17th International Conference, TSD 2014*, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655, 257-264. Springer International Publishing Switzerland.

Benko V. 2014b. Compatible Sketch Grammars for Comparable Corpora. In A. Abel, C. Vettori and N. Ralli (Eds.), *Proceedings of the XVI EURALEX International Congress: The User In Focus*. 15–19 July 2014, 417-430. Bolzano/Bozen: Eurac Research.

Čermák F. 1997. Czech National Corpus: A Case in Many Contexts. *International Journal of Corpus Linguistics*, 2(2): 181–197.

Čermák F. 2009. Spoken Corpora Design: Their Constitutive Parameters. *International Journal of Corpus Linguistics*, 14(1): 113–123.

Ebeling J. & Ebeling S. O. 2013. *Patterns in Contrast*. John Benjamins, Amsterdam.

Granger S. 2003. The Corpus Approach: A common way forward for contrastive linguistics and translation studies? In S. Granger, J. Lerot and S. Petch-Tyson (Eds.), *Corpus-based Approaches to Contrastive Linguistics*, 17–29. Rodopi, Amsterdam.

Greenbaum S. 1996. *Comparing English World-Wide*. Clarendon Press, Oxford.

Hundt M. and Gut U. (Eds.). 2012. *Mapping Unity and Diversity World-Wide*. John Benjamins, Amsterdam.

Johansson S. 2007. *Seeing through Multilingual Corpora: On the use of corpora in contrastive studies.* John Benjamins, Amsterdam.

König E. and Gast V. 2012. *Understanding English-German Contrasts*. Erich Schmidt Verlag, Berlin.

Mauranen A. 1999. Will 'translationese' ruin a contrastive study? *Languages in Contrast,* 2 (2): 161-185.

[6] http://universaldependencies.org/ and http://unesco.uniba.sk/aranea_about/index.html (Benko, 2014a, b)

[7] Korpusanalyseplattform der nächsten Generation; cf. http://www1.ids-mannheim.de/kl/projekte/korap.html

[8] By CC BY-NC licensing; f. https://creativecommons.org/licenses/by-nc/2.0/

Marzo S., Heylen K. and De Sutter G. (Eds.). 2012. *Corpus Studies in Contrastive Linguistics*. John Benjamins, Amsterdam.

McEnery T. and Hardie A. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, Cambridge.

McEnery T. and Xiao R. 2007. Parallel and Comparable Corpora: What is Happening? In G. M. Anderman and M. Rogers (Eds.), *Incorporating Corpora: The Linguist and the Translator,* 18–31. Multilingual Matters, Clevedon.

O'Keeffe A. and McCarthy M. 2010. *The Routledge Handbook of Corpus Linguistics*. Routledge, Abingdon.

O'Keeffe A., McCarthy M. and Carter R. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge University Press , Cambridge.

Ostler N. 2008. Corpora of less studied languages. In A. Lüdeling and M. Kytö (Eds.), *Corpus Linguistics: An International Handbook,* 457–483. Mouton de Gruyter, Berlin.

Sharoff S., Rapp R., Zweigenbaum P. and Fung P. (Eds.). 2013. *Building and Using Comparable Corpora*. Springer, Heidelberg.

World Englishes (1996) vol. 15(1), special issue on the International Corpus of English, guest-edited by S. Greenbaum and G. Nelson.

World Englishes (2017) vol. 36(3), special issue on the International Corpus of English, guest-edited by G. Nelson, R. Fuchs and U. Gut.

Xiao R. 2008. Existing Corpora. In A. Lüdeling and M. Kytö (Eds.), *Corpus Linguistics: An International Handbook,* 383–456. Mouton de Gruyter, Berlin.

Creating CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes - The National Corpus of Contemporary Welsh)

Dawn Knight[1], **Tess Fitzpatrick**[2], **Steve Morris**[2], **Jeremy Evas**[1], **Paul Rayson**[3], **Irena Spasić**[1], **Mark Stonelake**[2], **Enlli Môn Thomas**[4], **Steven Neale**[1], **Jennifer Needs**[2], **Scott Piao**[3], **Mair Rees**[2], **Gareth Watkins**[1], **Laurence Anthony**[4], **Thomas Michael Cobb**[5], **Margaret Deuchar**[6], **Kevin Donnelly**[7], **Michael McCarthy**[8], **Kevin Scannell**[9]
[1]Cardiff University, [2]Swansea University, [3]Lancaster University, [4]Bangor University, [4]Waseda University, [5]University of Quebec at Montreal, [6]University of Cambridge, [7]Freelance, [8]University of Nottingham, [9]Saint Louis University

CorCenCC is an inter-disciplinary and multi-institutional project that is creating a large-scale, open-source corpus of contemporary Welsh. CorCenCC will be the first ever large-scale corpus to represent spoken, written and electronically-mediated Welsh (compiling an initial data set of 10 million Welsh words), with a functional design informed, from the outset, by representatives of all anticipated academic and community user groups.

The CorCenCC project is led by Cardiff University with academic partners at Swansea, Lancaster and Bangor Universities. It has received major funding of £1.8M from two UK research councils (ESRC and AHRC) and attracted contributions and support from stakeholders including the Welsh Government, National Assembly for Wales, BBC, S4C, WJEC, Welsh for Adults, Gwasg y Lolfa, and University of Wales Dictionary of the Welsh Language. Nia Parry (TV presenter, producer and researcher; Welsh tutor, Welsh in a week (S4C)); Nigel Owens (international rugby referee; TV presenter), Cerys Matthews (Musician author; radio and TV presenter) and Damian Walford Davies (Prof. of English Literature; poet Chair of Literature Wales) are the official ambassadors of the CorCenCC project which started in March 2016, and lasts for 3.5 years.

The corpus will enable, for example, community users to investigate dialect variation or idiosyncrasies of their own language use; professional users to profile texts for readability or develop digital language tools; to learn from real life models of Welsh; and researchers to investigate patterns of language use and change. Corpus design and construction in a minority language context such as that of Welsh poses interesting challenges, but also presents opportunities perhaps not open to developers of corpora for larger languages.

In our presentation, we provide an overview of the whole project highlighting key elements such as:

- Collection, transcription and anonymisation of the data: so far, we have extended our initial plans and developed a sampling frame for the corpus
- Development of the part-of-speech tagset and tagger: including ongoing work to create a gold-standard data for training and evaluating the Welsh natural language processing tools
- Development of a semantic annotation tool: the project has adapted the UCREL Semantic Analysis System (USAS) taxonomy for Welsh and a prototype semantic tagger has been created
- Scoping and construction of an online pedagogic toolkit: to date we have undertaken surveys with stakeholders, national and international advisors in order to collect requirements for this tool
- Infrastructure to collect and host the resulting corpus: this involves designing and building a crowdsourcing app (currently available for iOS with Android under development) for the general population to donate their conversational data, alongside the design of storage and retrieval software

Four presentations at the main CL2017 conference (Rees et al., 2017; Piao et al., 2017; Needs et al., 2017; Neale et al., 2017) provide more detail on these aspects. Further details of the project are available from the website: http://www.corcencc.org/

Acknowledgments

CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes The National Corpus of Contemporary Welsh): A community driven approach to linguistic corpus construction is an interdisciplinary,

collaborative project led by the School of English, Communication and Philosophy at Cardiff University. The project commenced on 1st March 2016 and is funded by the Economic and Social Research Council (ESRC) and the Arts and Humanities Research Council (AHRC) (ref. ES/M011348/1).

References

Rees, M., Watkins, G., Needs, J., Morris, S. and Knight, D. 2017. Creating a Bespoke Corpus Sampling Frame for a Minoritised Language: CorCenCC, the National Corpus of Contemporary Welsh. *Paper presented at the CL2017 conference*, University of Birmingham, Birmingham, 24-28 July 2017.

Piao, S., Rayson, P., Knight, D., Watkins, G. and Donnelly, K. 2017. Towards a Welsh Semantic Tagger: Creating Lexicons for A Resource Poor Language. *Paper presented at the CL2017 conference*, University of Birmingham, Birmingham, 24-28 July 2017.

Needs, J., Knight, D., Morris, S., Fitzpatrick, T., Thomas, E. and Neale, S. 2017. "How will you make sure the material is suitable for children?": User-informed design of Welsh corpus-based learning/teaching tools. *Paper presented at the CL2017 conference*, University of Birmingham, Birmingham, 24-28 July 2017.

Neale, S., Spasi, I., Needs, J., Watkins, G., Morris, S., Fitzpatrick, T., Marshall, L. and Knight, D. 2017. The CorCenCC Crowdsourcing App: A Bespoke Tool for the User-Driven Creation of the National Corpus of Contemporary Welsh. *Paper presented at the CL2017 conference*, University of Birmingham, Birmingham, 24-28 July 2017.

EuReCo - Joining Forces for a European Reference Corpus as a sustainable base for cross-linguistic research

Marc Kupietz[1], Andreas Witt [1,2,3], Piotr Bański[1], Dan Tufiş[4], Dan Cristea[5,6], Tamás Váradi[7]

[1] Institut für Deutsche Sprache, Mannheim
[2] University of Cologne, Faculty of Arts and Humanities
[3] Heidelberg University, Department of Computational Linguistics
[4] Institute for Artificial Intelligence Mihai Drăgănescu, Bucharest
[5] Romanian Academy, Institute for Computer Science - Iaşi
[6] "Alexandru Ioan Cuza" University of Iaşi, Department of Computer Science
[7] Research Institute for Linguistics, Hungarian Academy of Sciences
(kupietz|witt|banski)@ids-mannheim.de, tufis@racai.ro, dcristea@info.uaic.ro, varadi.tamas@nytud.mta.hu

Abstract

In this paper we discuss the opportunities, prerequisites, possible applications and implications of a virtually joint corpus based on existing national, reference or other large corpora and their host institutions.

1 Introduction

The past 20 years have seen an emergence of national, reference and other large corpora of numerous European languages (Aston & Burnard, 1998; Váradi, 2002; CNC, 2005; Geyken, 2007; Baroni et al., 2009; Davies, 2010; Kupietz et al., 2010; Przepiórkowski et al., 2010; Oravecz et al., 2014; Tufiş et al., 2016). Most of them have been or are being built in projects of limited duration, but typically based at institutions that are at least to some degree responsible for curating data and for making it available to the respective scientific communities also after the building phase. The idea of EuReCo, which has been around in the CMLC workshop series since 2012 (see Bański et al., 2012), is that such institutions, rather than continuing as "research islands", should join forces and experiment whether a well-designed technology could allow a unifying view on building and exploitation of a multilingual collection of comparable corpora, a goal motivated by the rapidly changing and growing variety of needs of the linguistic and related user communities.

We present in this paper such a joint project, called EuReCo, briefly showing its aims, the technology behind and the language resources involved.

2 Aims

2.1 Comparable corpora

One of the aforementioned growing needs is the need for comparable corpora in order to facilitate contrastive and generally cross-linguistic research beyond the possibilities provided by parallel corpora, which are very much limited for linguistic applications by unavoidable translation biases. This application area is also the initial and currently the main focus of EuReCo. It appears that joining forces in this area is a particularly promising prospect: given that several national and reference corpora are built and maintained anyway and independently, with methodologies and techniques developed for joining them virtually, where each national centre is still responsible for its language and each corpus still physically located at its centre, it should be much more economical, scalable and sustainable to build a single virtual comparable corpus linking these existing resources than to create the comparable corpora from scratch, possibly even at more than one centre.

2.2 Further aims

In the meantime, however, the envisioned EuReCo has acquired a broader range of potential applications: if the organisational and technical prerequisites for such an infrastructure prove feasible, it would be wise to identify – as early as possible – further functionalities that are currently required or envisioned by the collaboration partners, such as, for example: the ability to manage very large corpora, statistical analysis – ideally dynamically offered to the user, or support for querying different kinds of linguistic annotations.

The general goal of the EuReCo initiative is to bring together existing European corpus initiatives,

specifically in those areas where synergy effects can be expected with high certainty and in a very much target-oriented fashion, towards goals that the collaboration partners would like to achieve, but are unlikely to achieve alone in a sufficiently effective and sustainable way.

Apart from these rather economical aspects, EuReCo also expects benefits from bringing closer together research communities that are currently centered around philologies and their sub-disciplines.

2.3 Relation to CLARIN

The EuReCo objectives are much narrower and oriented towards target applications than those of the European Language Resource Infrastructure Project CLARIN, which "*makes digital language resources available to scholars, researchers, students and citizen-scientists from all disciplines, especially in the humanities and social sciences, through single sign-on access.*" (see www.clarin.eu). In contrast to CLARIN, which has been particularly strong at providing horizontal base layers of infrastructure, standards and best practices, EuReCo will typically aim at vertical columns ending directly at end-user applications.

3 Foundations

Despite the differing scope and objectives, EuReCo will necessarily be tightly integrated into CLARIN. In addition, its roots lie in a number of experiences gathered by its collaboration partners and their respective histories of providing corpora and tools for using them, in large part within CLARIN:

- contemporary corpora are always tied to their hosting organizations by license contracts and other legal restrictions (Kupietz *et al.*, 2014),
- the way linguists use corpus data is itself subject to rapidly developing research,
- the exact requirements of corpus search and analysis tools for different corpora differ with research traditions and target communities,
- there will be no single tool that satisfies all user needs,
- unification is often necessary to keep costs manageable and to allow for re-usability, but one has to be very careful to keep the results usable and useful.

Based on these insights, the EuReCo strategy can be characterized by the following key properties:

- the aims of the collaboration have to be carefully picked and outlined in order to guarantee that:
  - the product of the collaboration is actually useful for the collaboration partners and their research communities,
  - the overhead of the collaboration does not outweigh its synergy effects,
- commonly developed and used tools must acknowledge the fact that the corpus data itself may not leave its hosting organizations,
- the collaboratively developed tools will usually not replace, but only complement those already existing.

4 Previous and current work

4.1 KorAP

The current main technical basis for EuReCo is the corpus query and analysis platform KorAP that has recently been developed at the IDS (Bański *et al.*, 2013; 2014, Diewald *et al.*, 2016). KorAP is the designated successor of the corpus search and management system COSMAS, which was launched in 1994 and in its second incarnation (COSMAS II), is still currently used by 39.000 researchers working on the German language. Besides KorAP's more performance-oriented features, such as horizontal scalability with respect to an unbounded corpus size and any number of annotation layers, two are particularly fundamental for EuReCo: (i) its ability to manage corpora that are physically located at different places, in order to comply with typical license restrictions (cf. Kupietz *et al.*, 2014) and (ii) its ability to dynamically create virtual sub-corpora based on text properties and to manage these virtual corpora in a persistent way, to e. g. allow for reusability and reproducibility. In addition, using a micro-service-like architecture, KorAP has been specifically designed for collaborative development and particularly collaborative extensibility up to the end-user. Extensibility is also KorAP's main approach to Jim Gray's famous postulate "*put the computation near the data",* which is essential not only to cope with big data, but also to cope with intellectual property rights (IPR) restrictions.

4.2 CoRoLa

CoRoLa is a priority project of the Romanian Academy, carried on by the Institute of Artificial Intelligence "Mihai Drăgănescu" in Bucharest and

the Institute for Computer Science in Iaşi, both affiliated with Romanian Academy. When finalised (end of 2017), CoRoLa will be the largest corpus of Romanian contemporary language, including both written and spoken data. The distinctive aspect of the CoRoLa project (Tufiş *et al.*, 2016) is that all the data included into the reference corpus have cleared IPR, based on bilateral agreements between the developing institutions and the data providers. The migration of CoRoLa data to the new DRuKoLA environment (see below) assumes new encoding and indexing methods, mapping annotations, etc., so that the users could enjoy all the facilities of the KorAP query platform.

4.3 The Hungarian corpus

The Hungarian National Corpus is a balanced reference corpus intended to capture varieties of five selected major genres of present-day Hungarian, namely journalism, literature, (popular science), personal, and official language use. Its first version appeared in 2001 and it contained 187 million running words, morphologically annotated and tagged. The majority of the data were collected from electronic sources from within Hungary but the HNC also contains subcorpora representing Hungarian as a minority language spoken in the neighbouring countries. On the design and implementation of the first release of the corpus see Váradi (2002).

The HNC has recently been substantially upgraded and extended to gigaword size. This new release followed the original design of the corpus but the internal proportions of the genres have been changed, mainly to do justice to the ubiquitous social media. The annotation has also been overhauled and the engine and the user interface have also been modernised, employing the Manatee/Bonito framework (Rychlý, 2007). Oravecz et al. (2014) describe the corpus in more detail.

4.4 DRuKoLA

Parts of the EuReCo vision have already been implemented in the DRuKoLA-project[1], large parts of which can also be regarded as a pilot study for EuReCo (Cosma *et al.*, 2016). DRuKoLA is centered around the German Reference Corpus DeReKo (Kupietz, *et al.*, 2010) and the Reference Corpus of Contemporary Romanian Language CoRoLa (Tufiş, *et al.*, 2015). One of its main objectives is to provide a common platform for constructing various kinds of comparable corpora, based on text properties and for analysing them for contrastive linguistic purposes.

The present state of the part of DRuKoLA relevant to EuReCo is that a converter from CoRoLa-TEI-format to KorAP-XML-format has been implemented so that CoRoLa can now be accessed via KorAP. For the present moment, a large part (60%, ˜300 million words) of the textual content of CoRoLa has been incorporated as the Romanian part of the DRuKoLA content. The next step will be to fine-tune a first version of mapping functions from CoRoLa and DeReKo metadata categories to intermediate taxonomies on the basis of which virtual corpora will be dynamically generated. It seems that intermediate taxonomies for topic domains and text types will typically be necessary to arrive at sufficiently valid and fine-grained common category systems.

Romanian speech data collected in CoRoLa will be added to DRuKoLa when the appropriate processing functionality of KorAP is finalized.

4.5 DeutUng

As a second EuReCo pilot project, *DeutUng*[2] will start to integrate the Hungarian National Corpus (HNC) into EuReCo. With respect to the establishment of an infrastructure and research methodology for comparable corpora, DeutUng is similar to DRuKoLA.[3]

5 Conclusions

The EuReCo initiative represents an ambitious effort of building a self-sustainable and flexible basis for comparable corpora, which is expected to offer very attractive opportunities for users but also challenges for developers. Multilinguality, which is at the root of the idea of EuReCo, together with

[1] DRuKoLA (2016-2019) is funded by the Alexander von Humboldt-Foundation, as a Research Group Linkage Programme, between the University of Bucharest and the Institute for the German Language in Mannheim, with the Institute for Artificial Intelligence *Mihai Drăgănescu* (RACAI, Bucharest) and the Institute for Computer Science (IIT, Iaşi) of the Romanian Academy as associated partners. The acronym combines central goals of the project: corpus development and contrastive linguistic analysis (*Sprachvergleich korpustechnologisch. Deutsch - Rumänisch*).

[2] DeutUng (2017-2020) is a co-operation project between IDS Mannheim and the University of Szeged with the Research Institute for Linguistics at the Hungarian Academy of Sciences as associated partner. It is also funded by the Alexander von Humboldt-Foundation as a Research Group Linkage Programme.

[3] With respect to linguistic application, however, DeutUng has as an additional focus on second language acquisition.

the vast repositories of language data, require innovative and robust technical solutions. The cooperation of several institutions and expert groups, as envisaged by EuReCo, promises to open new research avenues in the European digital humanities. Moreover, the technical base developed in EuReCo will provide support for innovative experiments that involve linguistic resources of different types and their interconnection. Showing that a commonly agreed methodology can provide unified access to very diverse basic level linguistic representations could provide useful insights concerning linking diverse types of linguistic data (corpora, dictionaries, wordnets, etc.) and unifying access to them.

6 References

Aston, G. and Burnard, L. (1998). The BNC Handbook: Exploring the British National Corpus with SARA. Edinburgh: Edinburgh University Press.

Bański, P., Kupietz, M., Witt, A., Ćavar, D., Heiden, S., Aristar, A. and H. Aristar-Dry (eds.) (2012): *Proceedings of the LREC-2012 workshop on "Challenges in the management of large corpora" (CMLC-1)*. Istanbul / Paris: ELRA.

Bański, P., Bingel, J., Diewald, N., Frick, E., Hanl, M., Kupietz, M., Pęzik, P., Schnober, C. and Witt, A. (2013): KorAP: the new corpus analysis platform at IDS Mannheim. In: Vetulani, Z. and Uszkoreit, H. (eds.): Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 6th Language and Technology Conference. Poznań: Fundacja Uniwersytetu im. A., 2013: 586-587.

Bański, P., Diewald, N., Hanl, M., Kupietz, M. and A. Witt (2014). Access Control by Query Rewriting: the Case of KorAP. In: *Proceedings of the 9th conference on the Language Resources and Evaluation Conference* (LREC 2014), European Language Resources Association (ELRA), Reykjavik, Iceland, May 2014: 3817-3822.

Baroni, M., Bernardini, S., Ferraresi, A., and E. Zanchetta (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation* 3/2009: 209-226.

CNC (2005). Czech National Corpus – SYN2005. Institute of Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic.

Cosma, R., Cristea, D., Kupietz, M., Tufiş, D. and A. Witt (2016). DRuKoLA – Towards Contrastive German-Romanian Research based on Comparable Corpora. In: Bański, P., Barbaresi, A., Biber, H., Breiteneder, E., Clematide, S., Kupietz, M., Lüngen, H., and A. Witt: 4th Workshop on Challenges in the Management of Large Corpora. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož / Paris: ELRA: 28-32.

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. Lit Linguist Computing (2010) 25(4): 447-464.

Diewald, N., Hanl, M., Margaretha, E., Bingel, J., Kupietz, M., Bański, P. and A. Witt (2016). KorAP Architecture – Diving in the Deep Sea of Corpus Data. In: Calzolari, Nicoletta et al. (eds.): *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16), Portoroz / Paris: ELRA: 3586-3591.

Geyken, A. (2007): The DWDS corpus: A reference corpus for the German language of the 20th century. *Collocations and Idioms*, London: 23–40.

Kupietz, M., Belica, C., Keibel, H. and Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, N. et al. (eds.): *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*: 1848-1854

Kupietz, M., Lüngen, H., Bański, P. and Belica, C. (2014). Maximizing the Potential of Very Large Corpora. In: Kupietz, M., Biber, H., Lüngen, H., Bański, P., Breiteneder, E., Mörth, K., Witt, A., Takhsha, J. (eds.): *Proceedings of the LREC-2014-Workshop Challenges in the Management of Large Corpora (CMLC2)*. Reykjavik / Paris: ELRA: 1–6.

Oravecz, Cs., Váradi, T. and Sass, B. (2014) The Hungarian Gigaword Corpus. In: Calzolari, Nicoletta et al. (eds.): *Proceedings on the Ninth International Conference in Language Resources and Evaluation (LREC'14).* Reykjavik / Paris: ELRA: 1719–1723.

Przepiórkowski, A., Górski, R. L., Łaziński, M. and P. Pęzik (2010). Recent Developments in the National Corpus of Polish. In Calzolari, N. et al. (eds.): *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Paris: ELRA.

Rychlý, P. 2007. Manatee/bonito–a modular corpus manager. In: *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno, Czech Republic: Masaryk University: 65–70.

Tufiş, D., Barbu Mititelu, V., Irimia, E., Dumitrescu, Ş. D., Boroş, T., Teodorescu, N. H., Cristea, D., Scutelnicu, A., Bolea, C., Moruz, A. and L. Pistol (2015). CoRoLa Starts Blooming – An Update on the Reference Corpus of Contemporary Romanian Language. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora* (CMLC-3). Mannheim: IDS: 5-10.

Tufiş, D., Barbu Mititelu, V., Irimia, E., Dumitrescu, Ş., D., Boroş, T. (2016). The IPR-cleared Corpus of Contemporary Written and Spoken Romanian Language. In: Calzolari, Nicoletta et al. (eds.): *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16), Portoroz / Paris: ELRA.

Váradi, T. (2002). The Hungarian National Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas /Paris: ELRA: 385–389.

CMC Corpora in DEREKO

Harald Lüngen
Institut für Deutsche Sprache
R5, 6-13
D-68161 Mannheim
luengen@ids-mannheim.de

Marc Kupietz
Institut für Deutsche Sprache
R5, 6-13
D-68161 Mannheim
kupietz@ids-mannheim.de

Abstract

We introduce three types of corpora of computer-mediated communication that have recently been compiled at the Institute for the German Language or curated from an external project and included in DEREKO, the German Reference Corpus, namely Wikipedia (discussion) corpora, the Usenet news corpus, and the Dortmund Chat Corpus. The data and corpora have been converted to I5, the TEI customization to represent texts in DEREKO, and are researchable via the web-based IDS corpus research interfaces and in the case of Wikipedia and chat also downloadable from the IDS repository and download server, respectively.

1 Introduction

The German Reference corpus DEREKO was started at the Institute for the German Language (IDS) in 1964 and has been continually expanded since then. Currently it contains more than 31 billion tokens and comprises text types as diverse as newspaper text, specialised texts, fiction, speeches and debates, computer-mediated communication and many more.

Though the bulk of DEREKO has always consisted of newspaper/press corpora, we have made new acquisitions in all of the above mentioned genres in the last couple of years (cf. e.g. Kupietz & Lüngen, 2014). In this paper, we would like to introduce three corpora of computer-mediated communication (CMC) that have recently been compiled for DEREKO. CMC is an interesting type of genre that is increasingly used in research on many aspects of language, e.g. interaction, neologisms, or orthography. In the following, we present our Wikipedia corpora in more detail, and in a little less detail the Usenet news corpus, and the Dortmund chat corpus, which make up the three types of CMC corpora currently in DEREKO.

2 Wikipedia

2.1 History

Since the 2000s, Wikipedia corpora have been created in cooperation with the IDS grammar department and have been included in DEREKO. In the first conversion 2005, the German Wikipedia dump, which contains the texts in the WP "wikitext", format was converted to CES, (Corpus Encoding Standard, cf. Ide, 1998), which was used to encode all IDS corpus holdings at that time). Strictly speaking, only Wikipedia talk pages (discussions) constitute CMC, but this first conversion included only the encyclopedia articles.

The 2011 conversion then for the first time included all German talk pages besides the articles and was produced using a new XSLT-based conversion pipeline which converted the wikitext directly into IDS-XCES encoding (Bubenhofer et al., 2011). It was decided that from now on a new Wikipedia conversion for DEREKO should be produced every two years, while the older conversions should always remain a part off DEREKO to enable diachronic analyses and anyway to ensure replicability of analyses. The 2013 conversion was done using an enhanced converter that employed the Sweble parser (Dohrn & Riehle, 2011), which was deemed a more sustainable method for parsing the wikitext format. The parsed wikitext was then passed to XSLT to produce the new target format I5 for DEREKO (Margaretha & Lüngen, 2014). I5 is a continua-

tion of IDS-XCES molded as a TEI P5 customisation which includes new elements, esp. <posting> adopted from Beißwenger et al. (2012), to represent the macrostructure of CMC dialogues as found in the talk pages. The 2013 conversion was characterised 2014 in the DeReKo paper Kupietz & Lüngen (2014). The 2015 conversion is characterised in the following section. Figure 1 gives an overview of the Wikipedia subcorpora in DEREKO sizes over the four conversions.

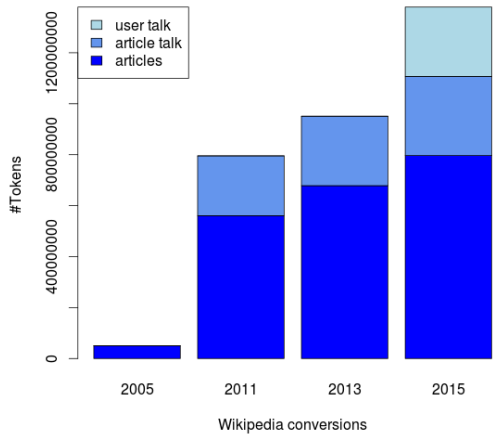

Figure 1: Sizes of WP subcorpora over the years

2.2 Latest WP Corpora

The following is an overview of the new features introduced in our 2015 Wikipedia conversion (which has been a part of DEREKO since 2016):

- User talk pages: Previously, Wikipedia corpora in DEREKO included only the article-related talk pages (where the WP editors discuss the structure and contents of the encyclopedia articles); since 2015 we also included the user talk pages. Every editor in Wikipedia can have a user talk page, and discussions can be conducted on these pages using the wiki software just like with article discussions. Here, users mainly discuss topics not related to the composition of the articles. For example, discussions on article talk pages that become off-topic are sometimes deferred to and continued on a user talk page. The user talk page corpus is of reasonable size similar to the article discussions corpus and constitutes a third Wikipedia corpus within DEREKO (see Fig. 1 and Table 1).
- Language Links: The metadata of the WP article corpus additionally contain all the language links of an article (links to WP pages for the same lemma in different languages). This is useful for creating comparable corpora from different language versions of Wikipedia for cross-lingual analyses. Even the talk pages with discussions about the articles of the same lemma in different languages can be related via these links.
- The I5 <autoSignature> element used in the representation of the discussions now comes with additional type information: "signed", "unsigned", or "user_contribution".
- Timestamps in discussions are now retained in the text (not only in the metadata) and are marked up using the I5 element <timestamp>). When researching WP discussions in COSMAS II, this is the only place where a user can see the date of a user edit.
- The wikitext to I5 converter has been improved, e.g. regarding timestamp identification, posting segmentation, thread identification, and the introduction of properties files for its configuration

	Articles	**Article talk pages**	**User talk pages**
I5 filesize	20G	5.5G	5.2G
#pages	1,802,682	591,460	539,053
#posts	--	6,200,701	5,523,769
#tokens	796,638,747	309,897,027	271,441,322

Table 1: Size of German Wikipedia corpora (conversion 2015) in DeReKo

2.3 Access

All of the above mentioned Wikipedia corpora (conversions from 2005, 2011, 2013, 2015) are included in DEREKO, the German Reference Corpus and can be researched via the COSMAS II, the Corpus Search, Management and Analysis System at the IDS. Presently, no POS annotations are provided for the Wikipedia corpora due to RAM limitations and the way COSMAS II handles annotation indexes. However, with the successor system KorAP (Bański et al., 2013;

	Wikipedia Corpora		
Lang.	**Articles #tokens**	**Article discussions #tokens**	**User discussions #tokens**
de.	796,638,747	309,897,027	271,441,322
en.	2,403,943,177	1,270,217,981	2,698,338,998
fr	764,459,026	131,107,729	372,639,260
hu.	117,987,947	8,293,799	26,215,158
no	99,014,144	5,314,362	32,481,331
es	578,883,431	54,907,258	276,034,367
hr	46,641,724	2,480,966	18,731,167
it	463,022,806	49,826,036	125,573,567
pl	298,207,197	16,558,557	64,126,136
ro	87,117,385	5,195,240	--

Table 2: Overview of sizes Wikipedia corpora in different languages. Interestingly, the German corpora are the only ones where the user discussion corpus is smaller than the articles discussion corpus, and the English corpora are the only ones where the user discussion corpus is bigger than the articles corpus.

Diewald et al., 2016), which has been in public beta test since May 2017, several linguistic annotation layers are already searchable. Besides querying the corpora in COSMAS II and KorAP, corpus, computational linguists can download the I5 files of all Wikipedia corpora from our pub server[1]. The wikitext to I5 converter (java jar files) can also be downloaded from there, complete with documentation (Margaretha, 2015).

2.4 Multilingual Wikipedia corpora

To prove that the conversion pipeline can be used for other Wikipedia language versions, we applied it to convert the Wikipedia dumps for nine further languages which play a role in contrastive and cross-lingual analysis projects at the IDS, see the overview in Table 2. Just like with the German WP, we generated the the corpus types WP articles, article discussions, and user discussions for each language in I5. They, too, are downloadable from our pub server and are even searchable in COSMAS II, where they reside in their own archive separate from DEREKO. However, since COSMAS II is a corpus interface designed for German language corpora, certain COSMAS functions cannot be meaningfully applied to the foreign language corpora, including tokenisation and lemmatization.

2.5 Related Work

The Berlin-based company *linguatools* offers monolingual and bilingual, comparable corpora built from the Wikipedia versions of 23 languages for free download. They are based on Wikipedia dumps from 2014 and contain the complete set of articles available in 2014, but only the articles, i.e. no talk pages. They contain rich metadata, including information about link types of internal and external links, and the WP categories under which an article is subsumed. They are distributed in XML markup and are downloadable from the company website (Linguatools, 2014). The bilingual corpora contain pairs of articles in language A and language B that were linked by Wikipedia language links. There are 23 monolingual and 253 bilingual comparable corpora available.

The linguatools Wikipedia corpus conversions cover more languages and contain somewhat richer metadata than ours. They do not include talk pages, and the XML encoding covers fewer structural phenomena than our I5 encoding. Their bilingual comparable WP-corpora are very useful for cross-lingual or contrastive linguistic analyses. Similar corpora could straightforwardly extracted from our Wikipedia corpora using the language links.

3 Usenet News

While many types of CMC corpora are alternatively identified as *social web corpora*, Usenet newsgroups definitely do not constitute a web genre, let alone a social web one, as the Usenet is based on its own internet protocol called nntp (Horton and Adams, 1987). As a CMC genre, newsgroups work similar to discussion forums, containing user contributions about a common topic organised in threads. Unlike typical Web 2.0 discussion forums, however, the Usenet is non-proprietary i.e. everybody can just use a news client and participate, or even set up their own news server to host newsgroups. Besides, all newsgroups are organised in a single topic hier-

[1] http://www.ids-mannheim.de/direktion/kl/projekte/korpora/verfuegbarkeit.html

archy, i.e. theoretically, for a particular topic, there is exactly one newsgroup. The Usenet started in 1979 and had its heydays in the 1990s, which makes it potentially interesting as a source of more historical, pre-Web 2.0 CMC.

We have compiled a corpus of German with all newsgroups from the news server news.individual.de (run by FU Berlin), with all textual newsgroups from the .de-hierarchy, starting in 2013. The downloaded news messages have been converted to I5 and been annotated with certain microstructural CMC features (Schröck & Lüngen 2015) and are researchable via COSMAS II, but for the time being (as the corpus is not anonymised) only on the premises of the IDS.

Usenet news corpus in DEREKO	
Period	2013-2016
#Newsgroups (all groups in the de. hierarchy)	375
#News messages	1,094,281
#Tokens	92,520,763

Table 3: Usenet news corpus overview

The news corpus in DEREKO is being continued with the latest data from the news server but also to be extended with data from the years before 2013. These, however, would have to be gleaned from a commercial news server.

4 Chat

In a so-called CLARIN-D curation project, the Dortmund Chat Corpus (about one million tokens, cf. Beißwenger et al., 2013) has been prepared for inclusion in CLARIN-D research infrastructures including DEREKO. The project work[2] comprised a conversion to a newly tailored CLARIN-D TEI customisation for chat and other CMC data (Lüngen et al., 2016), CMC-specific part-of speech tagging (Beißwenger et al, 2015), and corpus anonymisation according to the requirements set out in a legal expertise. The result is the Dortmunder Chat Corpus 2.0 as characterised in Table 4.

Dortmund Chat Corpus 2.0	
# log files	470
# posts	131,003
# tokens	1,005,166
File size (CLARIN-D TEI)	100 MB

Table 4: Dortmund Chat Corpus 2.0 Overview

It has been converted to I5 and is integrated in DEREKO and will be searchable through COSMAS II shortly. Besides, it is also available from the CLARIN-D repository at IDS[3].

5 Conclusion

We presented three types of CMC corpora that have recently been compiled at the IDS or curated from an external project and included in DEREKO, the German Reference Corpus, namely Wikipedia (discussion) corpora, the Usenet news corpus, and the Dortmund Chat Corpus. We will continue to build Wikipedia linguistic corpora every two years, i.e. the preparation of the 2017 conversion is impending. It will include a few new features, e.g. new metadata types similar to those available in the linguatools corpora, and also further types of discussion corpora from other Wikipedia namespaces. The Usenet corpora will be updated with the latest data but also be extended with data from the years before 2013. Chat corpora with more recent smart phone chat data will be acquired via a cooperation with the Mobile Communication Database project at the University of Duisburg-Essen.[4] We will also continue to try and compile other types of CMC corpora e.g. from web 2.0 blogs and discussions forums, provided that they come with licenses appropriate for redistribution.

References

Bański, P., Bingel, J., Diewald, N., Frick, E., Hanl, M., Kupietz, M., Pęzik, P., Schnober, C. and Witt, A. (2013): KorAP: the new corpus analysis platform at IDS Mannheim. In: Vetulani, Z. and Uszkoreit, H. (eds.): Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 6th Language and Technology Conference. Poznań: Fundacja Uniwersytetu im. A., 2013. Pages 586-587.

[2] together with partners from the Universities of Mannheim, Duisburg-Essen, and the Berlin-Brandenburg Academy of Sciences.

[3] PID: http://hdl.handle.net/10932/00-0379-FDFE-CC30-0301-E

[4] http://mocoda.spracheinteraktion.de/

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L., and Storrer, A. (2012). A TEI Schema for the Representation of Computer-mediated Communication. Journal of the Text Encoding Initiative, 3.

Beißwenger, M., Ehrhardt, E., Horbach, A., Lüngen, H., Steffen, D., Storrer, A. (2015): Adding Value to CMC Corpora: CLARINification and Part-of-speech Annotation of the Dortmund Chat Corpus. In: Beißwenger, Michael/Zesch, Torsten (Hg.): NLP4CMC 2015. 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media. Proceedings of the Workshop , September 29, 2015 University of Duisburg-Essen, pages 12-16, German Society for Computational Linguistics & Language Technology (GSCL).

Bubenhofer, N., Haupt, Stefanie, Schwimm, Horst (2011): A Comparable Corpus of the Wikipedia: From Wiki Syntax to POS Tagged XML. Hamburg Working Paper in Multilingualism, 96 B. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-51897

Diewald, N., Hanl, M., Margaretha, E., Bingel, J., Kupietz, M., Bański, P., Witt, A. (2016): KorAP Architecture – Diving in the Deep Sea of Corpus Data. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., (eds.): Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, pages 3586-3591, Paris: European Language Resources Association (ELRA).

Horton, M.; Adams, R. (1987): RFC-1036 Standard for Interchange of USENET Messages. Available online at: http://tools.ietf.org/html/rfc1036 .

Dohrn, H. and Riehle, D. (2011). Design and implementation of the Sweble Wikitext parser: unlocking the structured data of Wikipedia. In Proceedings of the 7th International Symposium on Wikis and Open Collaboration, WikiSym '11, pages 72–81, New York, NY, USA. ACM.

Ide, N. (1998). Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic

Corpora. In: Proceedings of the First International Language Resources and Evaluation

Conference (LREC), pages 463–470, Granada, Spain.

Ide, N., Bonhomme, P., and Romary, L. (2000). XCES: An XML-based standard for linguistic corpora. In: Proceedings of the Second Language Resources and Evaluation Conference (LREC), pages 825–830, Athens, Greece.

Kupietz, M. and Lüngen, H. (2014): Recent Developments in DeReKo. In: Calzolari, Nicoletta et al. (eds.): Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages. 2378–2385. European Language Resources Association (ELRA).

Linguatools (2014): Wikipedia Comparable Corpora. http://linguatools.org/tools/corpora/wikipedia-comparable-corpora/

Lüngen, H., Beißwenger, M., Ehrhardt, E., Herold, A., Storrer, A. (2016): Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. In: Dipper, Stefanie/Neubarth, Friedrich/Zinsmeister, Heike (eds.): Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016). Bochumer Linguistische Arbeitsberichte (BLA) 16, pages 156-164, Bochum.

Margaretha, E. and Lüngen, H. (2014): Building linguistic corpora from Wikipedia articles and discussions. In: *Journal for Language Technologie and Computational Linguistics (JLCL)* 2/2014

Margaretha, E. (2015): Documentation of the IDS Wikipedia Converter, 2015. http://corpora.ids-mannheim.de/pub/tools/2015/

Schröck, J. and /Lüngen, H. (2015): Building and Annotating a Corpus of German-Language Newsgroups. In: Beißwenger, Michael/Zesch, Torsten (eds.): NLP4CMC 2015. 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media. Proceedings of the Workshop , September 29, 2015 University of Duisburg-Essen, pages 17-22, German Society for Computational Linguistics & Language Technology (GSCL).

Organizing corpora at the Stanford Literary Lab

Balancing simplicity and flexibility in metadata management

David McClure, Mark Algee-Hewitt, Steele Douris, Erik Fredner, Hannah Walser
Stanford Literary Lab, Department of English, Stanford University
{dclure, malgeehe, sdouris, fredner, walser}@stanford.edu

Abstract

This article describes a series of ongoing efforts at the Stanford Literary Lab to manage a large collection of literary corpora (~40 billion words). This work is marked by a tension between two competing requirements – the corpora need to be merged together into higher-order collections that can be analyzed as units; but, at the same time, it's also necessary to preserve granular access to the original metadata and relational organization of each individual corpus. We describe a set of data management practices that try to accommodate both of these requirements – Apache Spark is used to index data as Parquet tables on an HPC cluster at Stanford. Crucially, the approach distinguishes between what we call "canonical" and "combined" corpora, a variation on the well-established notion of a "virtual corpus" (Kupietz et al., 2014; Jakubíek et al., 2014; van Uytvanck, 2010).

1 Introduction

The Literary Lab[1] is a research group in the English department at Stanford University that applies computational methods to the study of literature. The raw data behind the lab's research output is a collection of about 20 full-text corpora that contain many hundreds of thousands of novels, plays, poems, essays, pamphlets, letters, and newspaper articles spanning roughly from 1500 to 2000. These corpora come in all shapes and sizes – everything from small, ad-hoc collections of plain-text files assembled by hand for individual projects up to very large, professionally-curated collections purchased by the Stanford Library. [2] In total, these data sets comprise about 4 terabytes of raw data, and contain about 40 billion words of text.

This paper will describe a set of ongoing efforts at the Literary Lab to build a system for accessioning, organizing, and analyzing these corpora – the pipeline that runs from the raw data sets that come through the door to the final statistics, plots, and insights that appear in articles and pamphlets[3] published by the lab. This has proven to be a difficult and interesting problem because it involves navigating a set of overlapping (and at times conflicting) requirements.

2 Simplicity versus flexibility

The crux of this, in many ways, is a tension between competing desires for both simplicity and flexibility. On the one hand, we want to put everything in a single place – we want some kind of a unified data model that provides a simple, structured way to interact with the data, something that lends itself to the type of quick experimentation and hypothesis-testing that's needed in a research context. We don't want to rewrite the same ETL ("extract, transform, load") bindings onto the corpora over and over again for each project, and don't want to duplicate common preprocessing steps like tokenization, part-of-speech tagging, lemmatization, dependency parsing, etc. And, maybe most important, we often want a frictionless way to easily work *across* corpora. For example, in a study of structural changes in the American novel over the 19th and 20th centuries,

[1] http://litlab.stanford.edu/

[2] Among others – the Eighteenth and Nineteenth Century Collections Online and American Fiction corpora from Gale, the British Periodicals Online and Chadwyck Healey corpora from ProQuest, the Early English Books Online corpus from the Text Creation Partnership, and the Chicago Novel Corpus. As Tiepmar (2016) notes, there is very little standardization across text corpora used in the digital humanities, and these are no exception.

[3] https://litlab.stanford.edu/pamphlets/

we need a way to combine the ~18k novels in the Gale American Fiction corpus (1820-1940)[4] with the partially-overlapping ~10k novels in the Chicago Novel Corpus (1880-2000). We need a way to merge corpora, to bridge across them, to fuse them together into seamless collections of texts that can be easily analyzed as a unit.

But, at odds with this impulse to flatten everything out into a common data format, we also don't want to put any kind of inherent constraints on the types of questions that we could theoretically ask of the data. For example, we almost always want to preserve the domain-specific (and often very idiosyncratic) metadata that comes with the individual corpora. To give one small example – with the British Periodicals Online[5] corpus, a collection of 5 million articles from 1720-1940, recent projects have needed to make extensive use of the `<ObjectType>` element in the original XML, which classifies each article according to a custom vocabulary – "Fiction," "Review," "Advertisement," "Correspondence," "Poem," etc. At first we tried to pick a single, flexible metadata standard that could accommodate texts from all of the corpora and map these types of fields into this common schema. But this became unwieldy – taken together, the corpora have extremely heterogeneous metadata, and while it was possible to map everything into a single schema, we quickly ended up with a kind of Frankenstein format in which individual metadata fields become confusingly overdetermined and start to mean very different things in different contexts.

And, in some cases, corpora come with data that is almost impossible to fit into any kind of standardized schema designed for books or articles. For example, a graduate student in the Lab is working with a corpus of ~16k books scraped from a "fan fiction" website,[6] which includes metadata like the number of "favorites" and "follows" on the book, lists of characters, information about when individual chapters were published or updated, and even additional entity types like reviews and comments, all of which would be difficult to shoehorn into a one-size-fits-all schema. And yet, for that particular corpus, all of this information is extremely useful from a research standpoint.

3 Canonical and combined corpora

On the one hand, then, a desire to have everything in the same place; but, on the other hand, a practical need to retain a pristine copy of the original metadata for each corpus. Over the course of the last few months, we have been experimenting with a new workflow that tries to accommodate both of these requirements at once. The project is a Scala codebase that uses Spark[7] to write data as Parquet[8] tables on Stanford's Sherlock cluster[9], a 120-node HPC cluster administered by the Stanford Research Computing Center.

The key idea is to distinguish between what we call "canonical" and "combined" corpora. (This is similar to the notion of a "virtual" corpus described by Kupietz (2014), Jakubíek (2014), and van Uytvanck (2010).) Bindings for the "canonical" corpora wrap the raw, upstream data that comes from the vendor and extract exact copies of all included metadata, generally preserving the original nomenclature exactly to avoid any ambiguity about where a field came from – for example, a column in a CSV of authors called "Secondary Occupation" would become a `secondaryOccupation` field. Entities from the corpora – novels, poems, plays, authors, profiles, reviews – are represented as separate Scala case classes, which, combined with Spark's `Dataset` API, provide a typesafe way to represent the different schemas, which makes it easier to avoid errors down the line when fusing them together into unified collections. These "canonical" corpus readers also index the full-text content in a standardized way – in addition to storing the unmodified plain text for open-ended analysis, a stream of parsed tokens is also stored as an array in each Parquet row, with each token annotated with basic metadata – the original word form in the text, a part-of-speech tag assigned by OpenNLP[10], start and end character positions, and the "offset," a 0-1 value that represents the word's ratio position inside the text. (This corresponds to "Level 1" annotation under the rubric of the Corpus Query Lingua Franca, as described by Evert et al. (2015).)

These bindings onto the raw corpora produce full-fidelity, stable versions of each corpus for use in Lab research. For projects that are just using

[4] http://www.gale.com/c/american-fiction-1774-1920

[5] http://www.proquest.com/products-services/british_periodicals.html

[6] https://www.fanfiction.net/

[7] https://spark.apache.org/

[8] https://parquet.apache.org/

[9] http://sherlock.stanford.edu/

[10] https://opennlp.apache.org/

one corpus, feature extraction jobs can be run directly against these canonical Parquet tables. No constraints are placed on the structure of these initial copies of the corpora – they can be exactly as simple or complex as needed to represent the raw transmission data.

Meanwhile, for projects that need to mix and match different corpora together – for example, a project that needs a unified novel corpus from Gale American Fiction, the British Library corpus, and the Chicago corpus – a new adapter is created that generates what we call a "combined" corpus, a temporary data set that is tailored around the needs of that specific project. The code to produce a combined corpus looks very similar to the code that wraps one of the original corpora, except that the combined corpus will simply read from the Parquet tables produced for each of the canonical corpora instead of directly parsing the raw transmission data. The combined corpus provides a custom metadata schema that merges together just the specific fields that are needed in the context of the project at hand, and the extraction job writes out a single Parquet table that serves as the "working" data set for that project. Last, in addition to defining this set of mappings by which the corpora are fused together for a project, the code that generates the combined corpus is also responsible for the key step of de-duplicating the texts that get mapped into the unified schema, using the MinHash / LSH approach described by Leskovec, Rajaraman, and Ullman in *Mining Massive Datasets* (2014). In the future, we plan to wrap this up as a structured API that can easily be reused when defining a new combined corpus.

Once these Parquet tables have been saved to disk – at which point there is no meaningful distinction between a canonical and combined corpus – these datasets can serve as the basis for the actual "analysis" or "query" jobs that ask specific research questions of the data. These jobs vary widely in size and scope. More often than not, they have more in common with what might be thought of as "feature extractors" than with "queries," in the strict sense of the idea – usually, instead of directly producing a result that can be interpreted by a researcher (eg, KWIC results), these analysis jobs will generate some kind of intermediate dataset tailored around a particular set of questions, generally small enough to fit in RAM on a regular computer – often CSV, JSON, or SQLite files, or binary models that have been trained on the corpora. These intermediate datasets are then usually moved off the HPC cluster and taken up in statistical environments like Jupyter notebooks or RStudio, where the final querying, analysis, and data visualization takes place.[11]

4 Problems and future directions

One thing we have struggled with is whether it makes sense to save the "combined" corpora to disk, or if they should be materialized on-the-fly when analysis jobs are run.[12] The downside to saving them, of course, is that we end up storing duplicate copies of the texts that get included in the combined corpora. For example, if three active projects use Gale American Fiction corpus, then we store it four times – once in the "canonical" table, and three times for each "combined" corpus. Jakubíek (2014) sees this as unworkable, and in some contexts it certainly would be – for example, in a public-facing project with hundreds or thousands of users, where duplicating portions of the corpora for each user would vastly increase the storage requirements.

But, in the context of an individual research group, this hasn't been a problem. The Lab has a 30-terabyte quota on the HPC cluster (of which we've never used more than 4-5), and the storage requirements for the combined corpora are more modest than they might seem. Because the combined corpora are inexpensive to generate – the computationally intensive work is done up-front by the canonical adapters – they can be treated as ephemeral data and deleted as soon as a project ends, making it unnecessary to store more than a handful at once.

Furthermore, from a standpoint of what might be thought of as "engineering ergonomics," there are some interesting advantages to saving the combined corpora as complete, self-contained pack-

[11] To pick up on Evert et al.'s taxonomy of "approaches" to querying corpora – most of the analysis jobs run by the Literary Lab fall into approach 3, where "requirements can only be satisfied by a Turing-complete query language." (Evert et al., 2015) Which, in this context, is just an open-ended Spark job written in Scala, Python, or R, operating on the raw or annotated text content.

[12] This could be accomplished fairly easily – a "join" table could be generated that would just store foreign-key references back to the texts in the canonical corpora along with the results of the deduplication process, and the texts could then be mapped together into a unified `Dataset` at runtime in Scala.

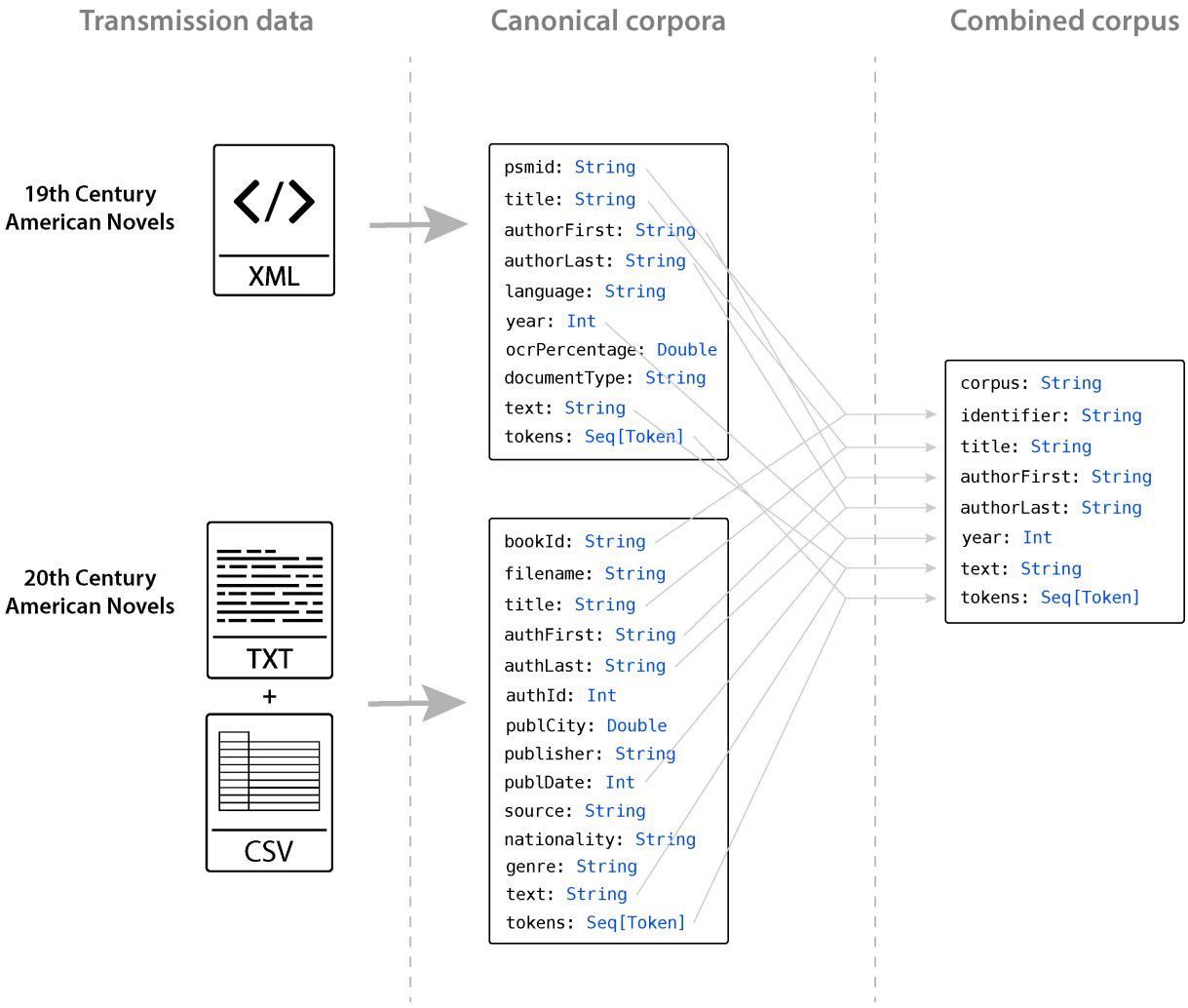

Figure 1: The data pipeline that produces a corpus of American novels spanning the 19th and 20th centuries. The transmission data for each input corpus is extracted into separate Parquet tables, keeping pristine copies of all the original metadata. These canonical representations of the corpora are then duplicated and merged together into a single "combined" corpus, which serves as the basis for the feature extraction jobs needed for the project at hand.

ages. For one thing it makes it very easy to share, publish, and archive the raw data that sits behind a particular project – the Parquet files can just be packaged up as a tarball and send to collaborators or dropped into an institution repository.

Even more important, though, at the level of research praxis – freezing off self-contained copies of the combined corpora makes it easy to maintain a separation between the code that *produces* the corpora and the code that *analyzes* the corpora. When analysis jobs can be written directly against static, unified datasets – instead of having to assemble input dataframes dynamically at runtime – they can easily be broken away from the corpus management system and structured as ad-hoc, decoupled, independent projects, which makes it easier for groups of researchers to iterate quickly on ideas without stepping on each other's toes in a single codebase. Meanwhile, the corpus management code itself can hew to the Unix philosophy of doing just one thing very well and focus exclusively on the task of accessioning and provisioning corpora, without getting cluttered up by analysis code. Research projects can be structured as sets of small, horizontally-scalable modules that interact with the self-contained datasets produced by the corpus manager.

That said, duplicating data in the combined corpora has obvious downsides, if only in that it puts a theoretical limit on the number of ways that we can mix and match the corpora, given a finite amount of storage. We're currently experimenting with hybrid models that would sidestep the need to duplicate the text data, while retaining the essential elements of this approach – the distinction between canonical and combined corpora, as well as the clean separation between corpus management and corpus analysis.

References

[Leskovec et al.2014] Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman. 2014. *Mining Massive Datasets*, 2nd edition (v2.1). Cambridge University Press, Cambridge, United Kingdom.

[Tiepmar2016] Jochen Tiepmar. 2016. CTS Text Miner: Text Mining Framework based on the Canonical Text Service Protocol. Proceedings of the 4th Workshop on Challenges in the Management of Large Corpora.

[Evert et al.2015] Stefan Evert and Andrew Hardie. 2015. Ziggurat: A new data model and indexing format for large annotated text corpora Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora.

[Kupietz et al.2014] Marc Kupietz, Harald Lüngen, Piotr Baski, and Cyril Belica 2014. Maximizing the Potential of Very Large Corpora: 50ăYears of Big Language Data at IDS Mannheim Proceedings of the 2nd Workshop on Challenges in the Management of Large Corpora.

[Jakubíek et al.2014] Milo Jakubíek, Adam Kilgarriff, and Pavel Rychlý 2014. Effective Corpus Virtualization Proceedings of the 2nd Workshop on Challenges in the Management of Large Corpora.

[van Uytvanck2010] Dieter van Uytvanck. 2010. CLARIN Short Guide on Virtual Collections. Technical report, CLARIN. http://www.clarin.eu/files/virtual_collections-CLARIN-ShortGuide.pdf

Accelerating Corpus Search Using Multiple Cores

Radoslav Rábara, Pavel Rychlý, Ondřej Herman and Miloš Jakubíček
Masaryk University
Botanická 68a
Brno, Czech Republic
{xrabara,pary,xherman1,jak}@fi.muni.cz

Abstract

The Manatee corpus management system on which the Sketch Engine is built is efficient, but unable to harness the power of today's multiprocessor machines. We describe a new, compatible implementation of Manatee which we develop in the Go language and report on the performance gains that we obtained.

1 Introduction

Text corpora are huge collections of texts in electronic form. They are used as an empirical resource for observation of real world language use, to study the behavior of words, their meanings and the contexts they occur in. Corpora are employed in many fields of linguistics (morphology, syntax, semantics, stylistics, sociolinguistics etc.) Important tools enabling corpus exploration are corpus managers. Corpus managers have to be able to deal with extremely large corpora effectively and provide platform for complex query evaluation, result filtering and visualization and computation of a wide range of lexical statistics. Processing speed is an important aspect of their operation because of the size of the corpora – billions of words and more. In order to speed up the processing, we reimplemented the single-threaded query evaluation engine in a concurrent way within the Manatee corpus management system (Rychlý, 2007).

2 Manatee system

The Manatee system (Rychlý, 2000) is a corpus manager, designed to be able to deal with extremely large corpora, optimized for fast query evaluation. It consists of an indexing library for text compression, index building and search, a query evaluation module, a query parser which transforms the textual query representation into abstract syntax trees, a set of command line tools for corpus building and maintenance and a graphical user interface. The system is based on the text indexing library FinLib which provides procedures for word indexing, corpus storage and retrieval of words in form of streams of positions (Rychlý, 2000). Manatee has its own query language, CQL, which enables users to execute complex queries on the corpus text.

The implementation of the query evaluation engine within Manatee is based on streams of tokens or token pairs, representing ranges or spans of consecutive tokens. The C++ `FastStream` and `RangeStream` interfaces represent token and range streams. Classes which implement them represent specific operations. The main idea is to have classes that perform simple operations which can be combined together to perform complex operations. In the original implementation, these classes are based on the iterator pattern. This means that only one value from the stream is available at any given time. The next value is loaded by calling the `next` method. Once it is called, the previous values are no longer available. Values are always provided in increasing order. After all values are read, an iterator returns a sentinel value, which is different from any other value in the stream. The iterator also provides a `find` method to seek to the next interesting value in an efficient way and a `peek` method to get the following value, which will be returned by calling the `next` method, without proceeding to the next position. More details about the implementation are in (Rychlý, 2000).

3 The Go programming language

Go, also referred to as Golang, is a new programming language which is being developed since 2007 by Google. Go tries to combine performance and security advantages of compiled language like C++ with the development speed of a dynamic language like Python (Pike, 2012). The language pro-

vides language-level parallelism through the so-called goroutines. A goroutine is a coroutine attached to a thread. Multiple coroutines can share a single thread to conserve operating system resources. The attachment is performed dynamically by the Go runtime. Communication and synchronization between goroutines is carried out using channels. Channels are used to move data from a sender to a receiver. The communication blocks the sender until the receiver receives the message. In this way, channels can provide synchronization between goroutines without explicit locks or condition variables. Channels can also be buffered. Buffered channel operations block the sender only when the buffer is full and they block the receiver only when the buffer is empty. More details about Go channels can be found in (Pike et al., 2012). The new implementation of Manatee is being developed in Go.

4 Implementation

In the new implementation, `FastStream` and `RangeStream` have been modified to provide a channel as the stream of the positions in place of the iteration protocol. The original methods `next`, `peek`, and `find` are no longer provided because their functionality is provided by the channel itself. The `find` operation was removed from the new implementation as it was, surprisingly, slowing the application down. In most cases we expected the `find` operation to improve the application performance by reducing the amount of the data transferred.

Not everything that the old implementation supports has been reimplemented and conversely, some functionality is not present in the old system, but the core functionalities of the old and the new systems are very similar.

We compared the complexity of a few selected modules which provide the same functionality by counting the lines of source code that are not blank or comment lines. Of the 4 compared modules, 3 of them (`FastStream` which represents sequences of positions, `Read_bits` and `Write_bits` which provide access to the compressed data storage, `SortedRuns` employed in lexicon construction) have nearly the same length in both of the implementations, the other module, `RangeStream`, used for used for handling sequences of structures or spans of text, is actually 30 % shorter in the new implementation, even though it supports concurrent query evaluation and employs some boilerplate to speed up communication over Go channels by sending larger batches.

5 Performance comparison

The evaluation of both of the implementations was performed on an eight core server. The original implementation is a single-thread application, so it is able to use only a single processor core. The new implementation, which has been designed in a concurrent way with goroutines, can exploit multiple cores. The new implementation was evaluated with different number of threads enabled. The performance was compared using a benchmark which measured the time of evaluation of a set of prepared queries. The prepared evaluation queries are complex and difficult to evaluate as they cover rules of syntactic analysis. The result of each of the queries is big, it covers approximately 5 % to 10 % of the whole corpus. The combined results of all the queries cover almost the whole corpus. These queries are quite extreme, but allow for more thorough evaluation of the system performance. Typical users of the Manatee system usually create simple queries to find specific words with a few restrictions, which usually produce small result sets.

The original implementation evaluated the prepared queries in 4h 29m 24s. The new implementation evaluated the queries in 2h 27m 39s, when running only on a single thread. Compared to the original implementation, the new implementation is faster by approximately 45 %. The results show that 13 of 15 of the queries were evaluated faster by the new implementation with an average speedup of approximately 45 %. The best improvement for a single query is 82 %. The smallest improvement is 20 %. Only two of the queries were evaluated slower by the new implementation. One was evaluated slower by 116 % and the second one was evaluated slower by 7 %.

The new implementation was evaluated with different amount of threads enabled, so that we could observe the performance scaling. As shown in Figure 1, the evaluation with three threads sped up the evaluation by 32 % compared to the evaluation with two threads and by 133 % compared to the evaluation with one thread. Four threads speed up the evaluation by 23 % compared to the evaluation with three threads and by 186 % comparing to the evaluation with one thread. Adding the fifth

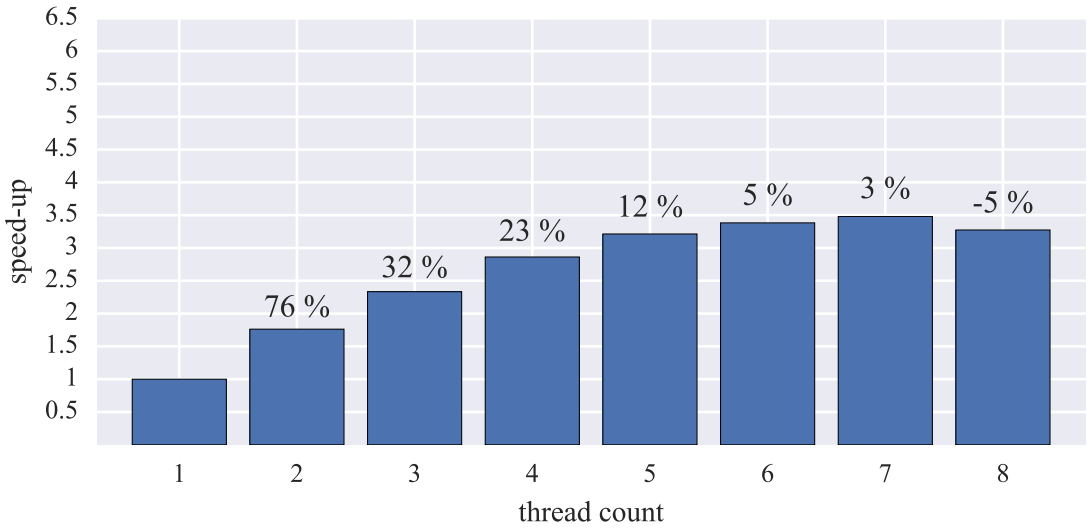

Figure 1: Average scaling over the whole testsuite

thread increases the performance by 12 % compared to the evaluation with four threads. Runs with six and seven threads differ in less than 5 % compared to runs with one thread. Using eight threads is actually slower than seven threads, likely due to I/O contention and cache spills.

6 Discussion

The speed-up observed with additional threads is more pronounced for complex queries, while simple queries might not scale at all. For example, a simple query of the form `[word="at"]` does not show any improvement in runtime when more threads are enabled, as can be seen in Figure 2. This is caused by the lack of the need for processing of the result. Most of the time needed to evaluate simple queries is spent on waiting for I/O and constructing the concordance.

A more complex query of the form `[word="at"][tag="NN"]` benefits from up to three threads. The query engine evaluates simultaneously the positions of the token *at* and positions of nouns. Another process then combines these two streams of positions and picks the pairs which represent positions that are next to each other.

The query[1] used in calculation of Word Sketches (Kilgariff et al., 2001) scales almost linearly when additional threads are employed, as can be seen in Figure 4. This is because the evaluation of the query needs to combine many different sources of data. I/O throughput is still important, but most of the time is spent in processes which manipulate the and combine the streams coming from storage.

7 Future work

While the new implementation of the query evaluation system is not significantly faster for simple queries, it provides large speed-ups for evaluating complex queries which are used for the calculation of Word Sketches, terminology extraction and other advanced features of Sketch Engine.

The new implementation is already used for calculation of some performance-intensive tasks, such as for the calculation of the Longest-commonest match (Kilgarriff et al., 2015), which was nearly infeasible with the old implementation.

Most importantly, the new architecture lays the groundwork for distributing the query processing over a larger cluster of machines, where every machine operates on a small part of the corpus only. This will allow us to provide further performance improvements, avoid I/O bottlenecks and also improve our ability to provide more redundant and fault tolerant system.

[1] `[word="it"] [tag="RB.?" | tag="RB" | tag="VM"]{0,3} [lemma="be" & tag="V.*"]? [tag="RB.?"]{0,2} [tag="DT.?" | tag="PP$"]{0,1} [tag="CD"]{0,2} [tag="JJ.?" | tag="RB.?" | word=","]{0,3} [tag="N.*"]{0,2} 1:[tag="N.*"] [word="for"] [tag="PP"] [tag="TO"]`

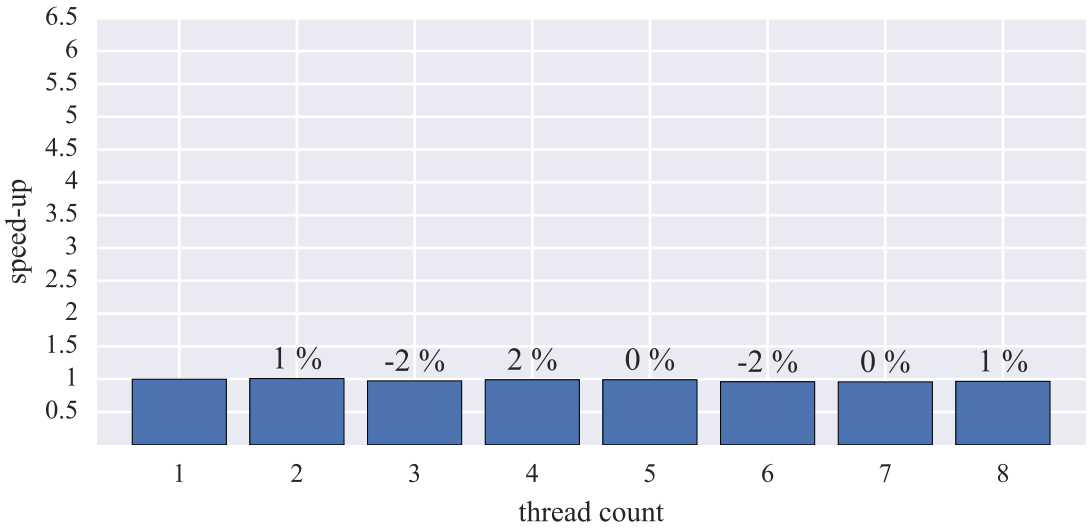

Figure 2: Scaling for a primitive query matching the word *at*

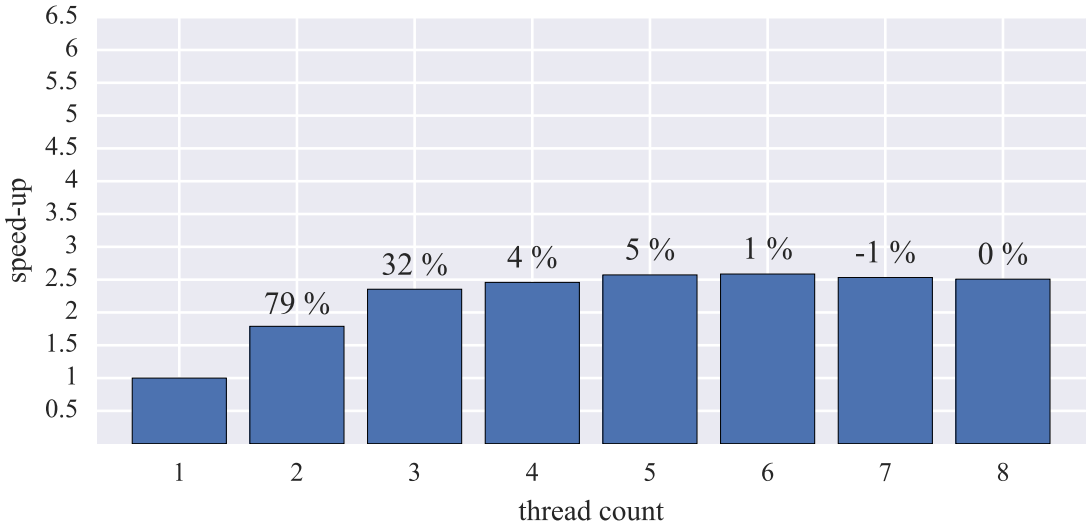

Figure 3: Scaling for a compound query matching sequences composed of *at* followed by a noun

8 Conclusion

The performance and the length of the source code were compared between the single-thread and concurrent implementation of the corpus manager Manatee. Manatee is able to deal with extremely large corpora and provides a platform for evaluating complex queries, filtering and visualizing results, and computing a wide range of lexical statistics (Kilgarriff et al., 2014). The original C++ implementation evaluated the prepared benchmark queries in approximately 4.5 hours. The new Go implementation managed to evaluate the prepared benchmark queries on a single thread in approximately 2.5 hours. The concurrent system performed better by 45 % on a single thread. The new implementation was also evaluated with different amount of enabled threads. The performance increased by 15.7 % on average with each additional enabled thread of the server and the most significant enhancement by 76 % was between running on one and two threads.

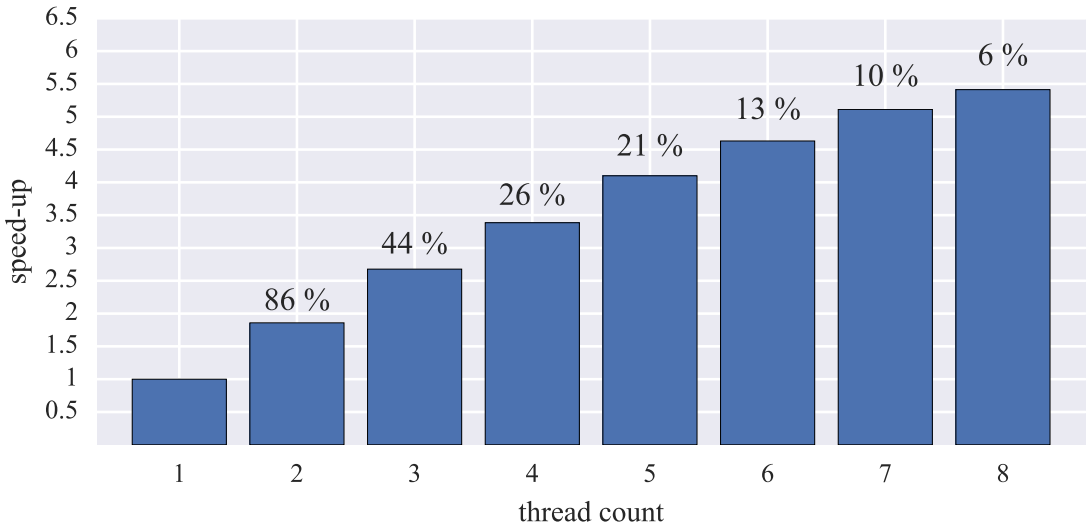

Figure 4: Scaling for a complex query matching phrases similar to *it's time for you to* and *it's not an intense thing for him to*

Acknowledgments

Reimplementation of Manatee in the Go language was partially supported by Lexical Computing.

References

Pike, R. 2012. Go at Google: Language Design in the Service of Software Engineering. online, URL: `https://talks.golang.org/2012/splash.article`

Pike, R., Gerrand, A. 2012. Concurrency is not parallelism. online, URL: `http://talks.golang.org/2012/waza.slide` Heroku Waza.

Rychlý, P. 2000. Corpus Managers and their effective implementation. PhD Thesis, Masaryk University.

Rychlý, P. 2007. Manatee/Bonito - A Modular Corpus Manager. *Proceedings of the 1st Workshop on Recent Advances in Slavonic Natural Language Processing*. pp.65–70. Masaryk University, Brno.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V. Lexicography, 1(1), pp.7-36. 2014. The Sketch Engine: ten years on. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Kilgarriff, A., Baisa, V., Rychlý, P. and Jakubíček, M. 2015. Longest–commonest Match. In *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference* (pp. 11-13).

Kilgarriff, A., Tugwell, D. 2001. Word sketch: Extraction and display of significant collocations for lexicography.

Keeping Properties with the Data
CL-MetaHeaders - An Open Specification

John Vidler
j.vidler@lancaster.ac.uk

Stephen Wattam
steve@watt.am

School of Computing and Communications
Lancaster University

Abstract

Corpus researchers, along with many other disciplines in science are being put under continual pressure to show accountability and reproducibility in their work. This is unsurprisingly difficult when the researcher is faced with a wide array of methods and tools through which to do their work; simply tracking the operations done can be problematic, especially when toolchains are often configured by the developers, but left largely as a black box to the user. Here we present a scheme for encoding this 'meta data' inside the corpus files themselves in a structured data format, along with a proof-of-concept tool to record the operations performed on a file.

1 Introduction

Corpora are continually increasing in size, and as a side effect, the management of this data during processing continues to be a pressing issue. As researchers we are under pressure to be able to reproduce the findings that we publish, and as the data we use increases in magnitude this can quickly become an onerous task. As noted in "*A Note on Rigour and Replicability*" (Louridas and Gousios, 2012), "*Unfolding off-the-shelf IR systems for reproducibility*" (Di Buccio et al., 2015) and "*Who Wrote the Web? Revisiting Influential Author Identification Research Applicable to Information Retrieval*" (Potthast et al., 2016), tracking the transforms performed on an input set is difficult at the best of times with the best of intentions, but when confronted with an unfamiliar tool or tool chain, inexperienced users can be forgiven for accidentally performing operations that do not achieve what they intend.

On the one hand, as a discipline, difficulties arising from the existence of a broad tool suite are an excellent problem to have; a multitude of solutions are present to handle any number of problems we may face. However, on the other hand, having no standard interoperability level for these tools prevents some interactions across this space without a great deal of effort from the user. Being aware of the file type is seldom enough to correctly set up reader/writer operations in a tool, especially if we consider character encoding differences, and the overall engineering of the tool has become very important in some cases. Further, processes for management of text metadata are tightly bound to the format, resulting in the need for bespoke tooling just to manage text flow through an NLP toolchain. An extensive overview of the issues has been compiled in "*A critical look at software tools in corpus linguistics*" (Anthony, 2013).

Here we propose a format for metadata storage that is sufficiently concise and flexible to be included *within* existing formats, allowing for metadata storage and processing in a manner minimally coupled with the underlying text storage regime.

Our approach draws upon the design principles of UNIX' 'magic numbers', most commonly encountered in BASH (Project, 2017) scripts. In shell scripts we see the "hash-bang" prefixed first line of each script which identifies the interpreter that should be used to read the file.

Such a hashbang line is a special case of a code comment as delimited by a single hash symbol `#`, normally reserved to be used to have the interpreter simply ignore the rest of a line. In the special "hash-bang" case, the first line of the script includes as `#!` with a system path string indicating which binary should be invoked to

execute the script. A common example of this is BASH interpreted scripts using `#!/bin/bash`.

Without this included in the header of the file, implementation differences between shell interpreters would quickly render scripts useless as the behaviour would be undefined. In older loaders without the initial "hash-bang" recognising capability, the loader can skip over the line as normal as the line starts with a comment symbol, effectively rendering it invisible, and preserving legacy behaviour.

We can borrow from this design for our own implementations, providing a mechanism for metadata storage and file type disambiguation that is flexible and simple to parse. Herein, we address the high-level design concerns surrounding such a format (Section 2) and the container format selected (Section 3) before going on to detail the mandatory fields designed to aid processing tools (Section 4). This paper is intended to introduce a preliminary specification, and further steps towards standardisation and use are discussed in Section 6 prior to an appendix containing examples for common NLP file formats.

2 Design

The primary aims of this specification are to provide a mechanism for detailing text- and toolchain-related metadata, and providing additional information on existing files that may affect subsequent processing stages.

Describing text metadata is a task already competently handled by formats such as TEI, and is primarily concerned with a rich, structured storage format that can be mined for information algorithmically. The need for such structure must be balanced here with the need for *compatibility* — the ability for the data to be 'hidden' from other NLP tools within their comment fields — and *universality* — the need for data to be applicable to many different types of toolchain.

In line with other bottom-up approaches, here we follow the design of assembling a minimal set of smaller, optional, features into a common data format. Following the lead of other minimal formats (such as vCards (Dawson and Howes, 1998) and JSON Schema[1]), we take a structured approach that assembles these components into a key-value map.

This approach permits the variation of representations for even the most basic data formats, providing they cover a basic subset. Here we require only 4 data types in accordance with the JSON specification(ECMA, 2013)[2], namely:

String Defined as the ASCII-7 subset only for reasons for character set compatibility;

Decimal Numbers Base-10, ASCII representation;

Nil Analogous to the empty set, `null`, `None` etc.

Boolean A single bit of information, i.e. `true` or `false`

These basic data types may be composed into data structures of two types: *objects*, which are key-value stores with any basic type as the key and another as the value, and *arrays*: objects with implicit, ordered integer keys. Keys are to be specified using `snake_case`. Keys starting and ending with two underscores '`__key__`' are reserved, using a convention similar to Python's PEP8(van Rossum et al., 2001).

With these as building blocks, we can construct structures which encompass text metadata, such as author, source, or date information, in addition to the possibility of describing sequences of operations - such as the tool history that has been performed to generate the file. In order to make the latter of these applicable to many toolchains, we specify a subset of structures that may be used to fulfil certain roles: this is intended to standardise and simplify tooling.

2.1 Namespaces and Interfaces

The basic metaheader structure is a single key-value map (*Object*) containing a set of keys defined by the current version of the specification (these are detailed later in Section 4). Each key within this top-level object points to a value, which may itself be a simple type or another Object. A field containing an object at the top level is said to provide a *namespace*.

Namespaces are specified separately to the core set of fields, and allow information to be grouped by purpose or administrator—this is similar to the approach taken by many packaging systems such as RubyGems[3] and PyPI[4]. Namespaces may be

[1] `http://json-schema.org/`

[2] And, it should be noted, many other data formats such as messagepack, BSON and YAML

[3] `https://rubygems.org/`

[4] `https://pypi.python.org/pypi`

defined by third parties in order to provide tool-specific key-value pairings — this mechanism is intended to allow tool authors to store information in formats that remain compatible with other tooling, without resorting to standoff annotation.

Because decentralised definition may lead to these namespaces becoming difficult to integrate and process, we reserve the capability to define a set of *interfaces*. Interfaces define a set of keys that must be provided together within a namespace, in order to offer a given service. This is analogous to duck typing in object-oriented languages: a namespace providing all of the fields required is presumed to behave according to the interface specification.

We reserve keys beginning and ending with two underscores ('`__key__`') for this purpose, for example, any namespace wishing to implement semantic versioning may provide a `__version__` field containing a semantic versioning compatible string. Any tooling wishing to support versioning of namespaces may then detect this and process such namespaces consistently.

3 Container Format

As opposed to simpler formats such as shell scripts, which need only know a single parameter to be able to select the correct interpreter at start-up, we can leverage modern structured data formats to embed effectively any amount of data into the header of a file. Many modern configuration tools use formats such as YAML(Evans, 2011) and JSON(ECMA, 2013), to provide rich options for holding data. YAML is commonly used, for example, as a metadata header section for static site generators (MkDocs, Jekyll and Hugo).

We should note that our intention is to specify the data structure contained within any such format, rather than the format itself, and propose that a number of formats may be implemented if necessary to remain within comment fields of other files.

The examples listed here, and the tooling that accompanies this paper, use JSON as a container format.

JSON was selected due to the breadth of its software support (according to `json.org` JSON is supported by 63 programming and scripting languages), and ease/speed of parsing. These properties make it suitable for use in many NLP contexts. Additionally, its simple data model and whitespace-agnostic form make it particularly suitable for representation within the comment fields of many NLP text formats.

JSON's popularity has lead to the existence of binary equivalent formats already available such as BSON(Group, 2015) and MessagePack(Furuhashi, 2013) for cases where data storage is in binary form.

3.1 Representation

As interoperability is key, the representation of data within the format is unspecified as far as possible. This means that the format may be included in a header or standoff documentation whilst remaining logic-compatible with all processing and aggregation tools. Further, inclusion in existing data formats is possible by simply adding the format to the comments.

Such conventions allow for multiple possible paths of information flow through NLP toolchains (an example of which is depicted in Figure 1): tooling that supports comments may retain the metaheaders in-place, or headers may be explicitly stripped out and processed in between text processing stages.

This approach mirrors that of the UNIX pipeline philosophy — small scripting tools may form the 'glue' around NLP toolchain components by reading headers and directing data as appropriate. This processing may be sufficiently generic to be handled by off-the-shelf tools (for example, the compilation of an audit trail including timestamps and processing arguments), or a custom processing stage using the metaheaders as a storage format.

This approach means that, for the first time, toolchain management code would be transferable throughout the community and between resources. In turn this enables the creation of interoperable tooling for documenting processing stages, aiding replicability.

4 Data Structure & Current Specification

Here we present an overview of the draft field specification designed to provide a minimal subset of fields to aid basic parsing and processing. Draft version '1.0.2' of the specification mandates no fields, but has 4

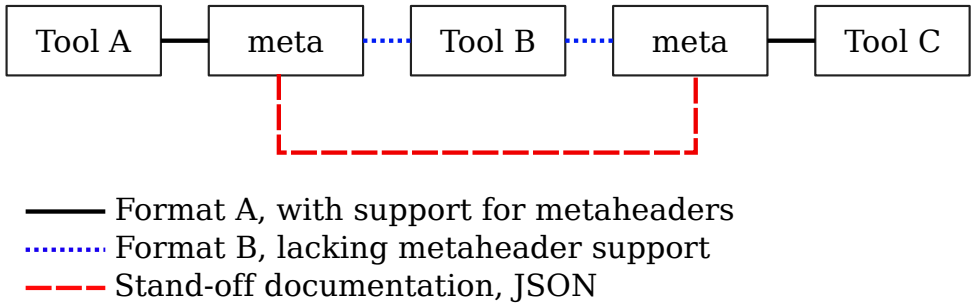

Figure 1: A sample workflow using the 'meta' tool to transfer annotations through a format that does not support commenting.

optional ones, some with default values, as described as follows:

`__version__` *(Optional, String)* - The specification version that this header complies with, if missing, assumes the latest draft. Version numbers are specified following Semantic Versioning (Preston-Werner,), making it easy to determine compatibility.

`encoding` *(Optional, String)* - Character encoding of current text, as defined by the IANA list of preferred text encoding names(Freed and Dürst, 2013). While optional, it is strongly advised that this field be present to avoid any ambiguity in parsing. If absent, this implicitly defaults to UTF-8.

`mime` *(Optional, String)* - The extended MIME(IANA, 2015) type used to describe what this file is.

`group` *(Optional, String or Object)* - Used to track which files belong to collections, defined as an object conforming to the `group` namespace specification.

`history` *(Optional, Object)* - A top-level namespace conforming to the `history` namespace specification that describes the processing history of this file. Complies with the interface specification and thus has an inner `__version__` field.

For the purposes of definition we draw the distinction between a *field*, where a key is assigned a simple value, and *namespace*, holding an object containing fields that themselves conform to a sub-format. Optional namespaces may then be assembled whilst retaining some guarantee of compatibility.

In the case of the above, `group` is specified as a namespace, which may contain:

`text_id` *(Required, String)* - A unique text identifier assigned to this text

`*` *(Optional, any)* - Further arbitrary key-value fields appropriate to the corpus

One simple example of where the non-string form of the `group` field could be used is that of parallel corpora; inner fields specifying a `collection` and `language` (See Figure 2) could be used to form the following structure to completely identify a part of the larger corpus.

```
...
"group": {
    "collection": "OPUS",
    "language": "en-gb"
},
...
```

Figure 2: An example of the `group` field being used as an object to identify this file as being part of the OPUS(`http://opus.lingfil.uu.se/`) UK English set.

The `history` namespace is intended to describe the history of a particular text's processing to form an audit trail of actions in the form of a list of actions. It is specified as:

`binary` *(Required, String)* - The program executed.

`time` *(Required, String)* - ISO8601 format datetime string describing the date and time at which the tool was run.

`args` *(Optional, String)* - The program arguments used.

platform *(Optional, String)* - The dot-delimited platform and architecture (ex. "Linux.x64").

md5 *(Optional, String)* - The md5 hash of the binary, used to ensure the correct version is used.

The set of features identified for inclusion in the first draft have been selected to allow the identification of key features of the subsequent texts, and allow them to be correctly loaded by software. They are common to all machine-readable text representations, describe necessary-yet-uninteresting features of the dataset, and are generally useful across many tool types.

These fields provide a level of process-accounting that so far has been absent from many NLP toolchains, and allows us to replay the processing that created the files in use.

In addition to the namespaces above, we define only a single interface designed to offer version reporting on third-party namespace specifications. Note that all fields are implicitly required for an interface to apply:

__**version**__ *(String)* - A semantic versioning compliant version string describing the version of the namespace that is being used.

4.1 Tooling

In addition to the format specification here, a proof-of-concept tool was developed, '`meta`' (source available at `http://ucrel.github.io/CL-metaheaders/`) which wraps existing commands and records their use in the files they generate, and can be used to validate existing meta headers in source files. Figure 3 shows how this command can be used to wrap existing tools to record their actions.

```
$> meta tagger \
    --input corpus.xml \
    --output result.xml
```

Figure 3: Running the '`meta`' tool on the program '`tagger`' with some arguments included. This would record the command in the `history` block inside `output.xml` metadata. Backslashes indicate a continued line.

Once the tool has been used to produce this history in the headers, the same tool can be used to extract the commands for later execution by the user.

The obvious use case for this is in cases where the user may have forgotten the precise commands they used, but is also useful for a second user to process other files in the same way as the first user, especially as part of a validation process.

Other included features of this tool are the ability to initialise a file with the basic metadata fields in a user-friendly way, and generate a list of dependencies for a given file. By reading the metadata of the file and getting the command history we can walk the list looking for any files that are required to generate the output given.

Furthermore, this can recurse through any recognised file that *also* has metaheaders included to create a full dependency tree which in turn can be used as part of the packaging process when files are to be distributed. This should aid researchers in producing correct source packages for distribution.

4.2 Extensions and Custom Namespaces

Further specifications and versions will be maintained and released in an open-source manner via the project's website at `http://ucrel.github.io/CL-metaheaders/`.

In addition to the specification (and tooling) provided here, we have designed the namespacing system to allow for other developers and tools to insert arbitrary data below the top level data structure. Proprietary fields are expected to use nested namespaces to keep the top-level clean, and allow developers the freedom to add their own variants for their own purposes - we do not expect to be able to predict all use cases for these headers.

Figure 4 demonstrates a TEI file with an additional 'software' sub-object to contain developer specific information (See Appendix 6 for further examples). The use of 'per-tool' namespaces in this manner allows for the use of standard file formats by various tools without loss of information that otherwise would have to be discarded after execution (or output in a difficult-to-track and proprietary standoff annotation format).

Because the software reading the files cannot know about all possible extensions to this format, we mandate that tools supporting the metadata specification must pass all unknown headers from

```
<?xml version="1.0"
   encoding="UTF-8"
   standalone="no" ?>
<TEI
  xmlns="http://www.tei-c.org/ns/1.0">
<!-- meta {
   "version": 1.0,
   "encoding": "UTF-8",
   "mime": "text/xml-tei",
   "software": {
      "author": "Joe Bloggs",
      "tool": "Jo's Awesome Software",
      "window": "+-5 words",
      "stoplist": true
   }
} -->
<teiHeader>
<fileDesc>
  ...
```

Figure 4: JSON data including an additional software description fields. Note that this is still version '1.0' compliant, as there is no restriction on additional data in the meta header. Newlines presented here are for the benefit of the reader, and can be entirely omitted for a single-line meta entry.

input to output without modification. They are, of course, free to change the fields and namespaces that account for any change applied to the text.

5 Further Work

Stated in Section 4, the standard is intended to provide only the barest minimum set to enable better communication between tools, and we fully expect to extend the format with additional data as future tools develop.

What we define here forms a 'core' field set, forming a number of reserved keys and associated namespace definitions. It is the authors' intent to allow developers the freedom to extend the standard with their own additions and as such we welcome any comments, suggestions regarding these extensions for inclusion in later standard releases.

One organisational addition is the creation of a registry of top-level namespaces. This will eliminate any potential issues with collision of third-party namespace definitions leading to incompatible implementations by offering a single versioned canonical list of namespace allocations.

In addition, we expect to produce further specification details for including meta data with archived corpus files, providing a mechanism for creating hierarchies of files.

Furthermore, as a living body of work, we intend to continue to integrate more document formats in to the standard as 'officially recognised types' and provide further examples of integration.

6 Summary

We have presented a simple method for including the properties of a file along with the file itself in a way that is backwards compatible with many existing text storage formats and tools. The basic design of this method is extensible in order to allow tool authors to annotate files, and to allow those building toolchains to use such data to manage existing tools. Using these capabilities, we present a proof-of-concept toolchain auditing application.

The specification outlined here is being actively developed, and a canonical reference (along with proof-of-concept and production tooling) are available at `http://ucrel.github.io/CL-metaheaders/`. Continued discussion of the specification, including bugs and feature requests can be done via the Github issues page at `https://github.com/UCREL/CL-metaheaders/issues`.

References

Laurence Anthony. 2013. A critical look at software tools in corpus linguistics. August.

F Dawson and T Howes. 1998. RFC 2426 - vCard MIME directory profile. `https://www.ietf.org/rfc/rfc2426.txt`, September.

Emanuele Di Buccio, Giorgio Maria Di Nunzio, Nicola Ferro, Donna Harman, Maria Maistro, and Gianmaria Silvello. 2015. Unfolding off-the-shelf IR systems for reproducibility. In *Proc. SIGIR Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR 2015)*.

ECMA. 2013. Ecma-404 - the json data interchange format. `http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf`, October.

Clark C. Evans. 2011. Yaml - yaml ain't markup language. http://yaml.org/, November.

Ned Freed and Martin Dürst. 2013. Iana, character sets. http://www.iana.org/assignments/character-sets/character-sets.xhtml, 12.

Sadayuki Furuhashi. 2013. Messagepack - it's like json, but fast and small. http://msgpack.org.

BSON Group. 2015. Bson - binary json. http://bsonspec.org/.

IANA. 2015. Media types. http://www.iana.org/assignments/media-types/media-types.xhtml, 10.

Jinho Choi, Universal Dependencies contributors. 2014. CoNLL-U Format. http://universaldependencies.org/format.html. Online; accessed June 2017.

Panos Louridas and Georgios Gousios. 2012. A note on rigour and replicability. *SIGSOFT Softw. Eng. Notes*, 37(5):1–4, September.

Martin Potthast, Sarah Braun, Tolga Buz, Fabian Duffhauss, Florian Friedrich, Jörg Marvin Gülzow, Jakob Köhler, Winfried Lötzsch, Fabian Müller, Maike Elisa Müller, Robert Paßmann, Bernhard Reinke, Lucas Rettenmeier, Thomas Rometsch, Timo Sommer, Michael Träger, Sebastian Wilhelm, Benno Stein, Efstathios Stamatatos, and Matthias Hagen, 2016. *Who Wrote the Web? Revisiting Influential Author Identification Research Applicable to Information Retrieval*, pages 393–407. Springer International Publishing, Cham.

Tom Preston-Werner. Semantic versioning 2.0.0. http://semver.org/.

The GNU Project. 2017. Bash - the gnu bourne-again shell. https://tiswww.case.edu/php/chet/bash/bashtop.html.

Guido van Rossum, Barry Warsaw, and Nick Coghlan. 2001. Pep 8: Style guide for python code. https://www.python.org/dev/peps/pep-0008/#id50.

A Further Examples For Common Formats

A.1 ARFF

```
% {"version":"1.0","encoding":"utf-8",\
  "mime":"application/x-weka"}
% 1. Title: Iris Plants Database
%
% 2. Sources:
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall
%      (c) Date: July, 1988
%
@RELATION iris

    @ATTRIBUTE sepallength   NUMERIC
    @ATTRIBUTE sepalwidth    NUMERIC
    @ATTRIBUTE petallength   NUMERIC
    @ATTRIBUTE petalwidth    NUMERIC
    @ATTRIBUTE class
{Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA ...
```

A.2 CoNLL-U

Note that implementations dealing with the CoNLL-U format are required to pass the contents of comments through their processing pipelines unaltered(Jinho Choi, Universal Dependencies contributors, 2014).

```
# {"version": "1.0","encoding":"utf-8",\
  "mime":"text/csv"}
# sent_id = 1
# text = Sue likes coffee and Bill \
  likes books
1       Sue         Sue
2       likes       like
3       coffee      coffee
4       and         and
5       Bill        Bill
...
```

Contributions of the Web as Corpus (WAC-XI) guest section

Organisers: Adrien Barbaresi (ICLTT Vienna), Felix Bildhauer (IDS Mannheim), Roland Schäfer (FU Berlin)

WAC-XI Homepage:

http://www.birmingham.ac.uk/research/activity/corpus/events/2017/cl2017/pre-conference-workshop-5.aspx

Are Web Corpora Inferior?
The Case of Czech and Slovak

Vladimír Benko
Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics
Panská 26, SK-81101 Bratislava
and
Comenius University in Bratislava
UNESCO Chair in Plurilingual and Multicultural Communication
Šafárikovo nám. 6, SK-81499 Bratislava
vladob@juls.savba.sk

Abstract

Our paper describes an experiment aimed to assessment of lexical coverage in web corpora in comparison with the traditional ones for two closely related Slavic languages from the lexicographers' perspective. The preliminary results show that web corpora should not be considered "inferior", but rather "different".

1 Introduction

During the last 15 years, creation of web corpora has been recognized as an effective way of obtaining language data in situations where building traditional corpora would be either too costly or too slow (Baroni et al., 2009; Jakubíček et al., 2013; Schäfer & Bildhauer, 2013) and building and analyzing web corpora has transformed into a separate branch of corpus linguistics.

At present, both traditional and web corpora do exist for many languages, with the respective web corpus being of comparable or even larger size. Any (corpus) linguist in this situation is therefore confronted with questions as follows: How does the existence of two "language samples" created by different methodology and technology influence my linguistic research? Which corpus provides better evidence allowing for generalizing my conclusions? Is any of the corpora "infereior"?

Both Czech and Slovak belong to languages where we can try looking for answers to such questions as respective corpora exist and the source data is (in our case) available.

2 Comparing Corpora

Due to the huge sizes of contemporary corpora, any comparison of their contents is a challenging task. For corpora available on-line, some comparisons can be performed via the respective interface, optionally in combination with the frequency lists generated from the respective corpora (Khokhlova, 2016). The large-scale statistical evaluation, however, requires having the source corpus data available (Kilgarriff, 2001).

Besides the assessment of lexical coverage based on rank and frequency distributions of word forms and/or lemmas, other corpus properties may also be compared, e.g. the "quality" of morphosyntactic annotation (out-of-vocabulary rate), "noise" (undetected foreign language and/or duplicate text fragments). If a tool for collocational analysis is available, such as Sketch Engine (Kigarriff et al., 2004; Kilgarriff et al., 2014), collocation profiles for a selected set of keywords can be conveniently compared.

3 The Experiment

In our paper, we describe an on-going experiment, in the framework of which we try to evaluate the lexical coverage of web corpora in comparison with the traditional corpora for the respective languages. As our comparison is mainly motivated by the needs of lexicographers, in an ideal case, it would be useful to compare the proportion of lexical items found in the respective corpora and not covered by existing dictionaries, that would qualify to become headwords in a newly complied dictionary (e.g., neologisms).

Such a task, however, would involve a lot of manual work – it is not enough just to count "out-of-vocabulary" tokens derived from the respective corpora: the web corpus naturally contains more of them because of more "noise".

We have therefore decided to do something that can be performed without any manual evaluation. The procedure involved comparing frequency lists derived from the respective corpora with headword lists of medium-sized dictionaries. As we were also interested how the corpus size influences the lexical coverage, we performed the same experiment with subcorpora of various sizes created by (random) sampling of the respective traditional and web corpus data.

3.1 The corpora

The traditional Czech corpora were represented by the *syn* series of the Czech National corpus (Křen et al., 2014) available from the LINDAT portal. The "opportunistic" *syn v4* basically contains all Czech corpus data gathered by the Institute of Czech National Corpus, making it rather unbalanced. A well-balanced part (containing the four representative 100 Megaword Czech corpora, i.e., *syn2000*, *syn2005*, *syn2010* and *syn2015,* respectively), however, can be easily extracted from *syn v4* by means of its metadata, yielding a balanced 400+ Megaword corpus that will be referred to as *syn20xx*.

The Slovak traditional corpora were represented by the *prim* series of the Slovak National Corpus (Šimková – Garabík, 2014; SNK, 2015). Two subcorpora have been used in our research – the 835 Megaword unbalanced *prim-6.1-all* (SNK, 2013a), and the 300+ Megaword balanced *prim-6.1-vyv* (SNK, 2013b). The source data of these corpora are, unfortunately, not available for users outside of our Institute.

The web corpora have been represented by the *Maximum* class of the Aranea Project corpora (Benko, 2014), i.e., the 5+ Gigaword *Araneum Bohemicum* for Czech, and the 3+ Gigaword *Araneum Slovacum* for Slovak.

To ensure the maximal compatibility of annotation among the corpora, both Czech and Slovak traditional corpora have been retokenized and retagged before being used in our experiment, which resulted in slight decrease of their original size measured in tokens. The information on corpora is summarized in Table 1.

Name	Language	Type	Size
syn 20xx	Czech	traditional, balanced	462 M tokens
syn v4	Czech	traditional	4,352 M tokens
Araneum Bohemicum Maximum (BM)	Czech	web	5,174 M tokens
prim-6.1-public-vyv	Slovak	traditional, balanced	317 M tokens
prim-6.1-public-all	Slovak	traditional	858 M tokens
Araneum Slovacum III Maximum (SM)	Slovak	web	3,357 M tokens

Table 1. Corpora used

3.2 Sampling Subcorpora

The subcorpora used in our experiment have been sampled in a logarithmic scale graded as follows: *1M*, *2M*, *5M*, *10M*, *20M*, …, etc., up to the actual corpus size. The rudimentary sampling algorithm was based on splitting each 1-Megaword block into two parts defined by the parameter. Though this procedure can be considered "radom" for very large subcorpora, it is certainly not the case with the small ones.

For each subcorpus, a frequency list has been extracted containing both lemmas and word forms, accompanied by the PoS information.

3.3 The wordlists

The only relatively new Czech dictionary available in electronic form that could be used to extract the Czech wordlist for our experiment was the (retro-digitized) bilingual Czech-Slovak Dictionary (Horák et al. 1981). The situation has been more favorable for Slovak, where several dictionaries in electronic form were available. We have opted here for the dictionary part the Rules of the Slovak Orthography (PSP, 2000), as its size is on par with the Czech dictionary used.

The extracted headword lists have been filtered to get rid of multi-word expressions (mostly secondary prepositions and loanwords), and to remove reflexive formants "se/si" for Czech and "sa/si" for Slovak that appear as parts of headwords with reflexive verbs, but would not have a counterpart in wordlists derived from corpora. After this processing the Czech list contained approx. 73,500, and the Slovak list 65,500 headwords, respectively.

.

Corpus size (M)	syn 20xx			syn v4			Araneum BM		
	(1+)	(10+)	(100+)	(1+)	(10+)	(100+)	(1+)	(10+)	(100+)
1	32.13	8.77	1.42	32.97	8.84	1.35	31.92	8.91	1.48
2	40.07	13.31	2.78	39.23	13.03	2.66	39.39	13.10	2.77
5	51.54	22.49	5.56	47.91	20.63	5.48	49.51	20.93	5.48
10	59.38	30.46	8.86	54.65	27.22	8.47	57.20	28.37	8.49
20	65.79	38.89	13.59	61.03	34.64	12.54	64.03	36.43	12.84
50	74.93	51.36	22.95	68.48	44.38	19.91	71.85	46.99	20.67
100	79.65	59.08	31.44	73.73	51.89	26.80	76.57	54.61	27.75
200	83.05	66.09	40.27	77.80	58.42	34.14	80.52	61.64	35.78
(<) 500	86.06	73.56	50.39	82.44	66.70	44.03	84.48	70.09	46.62
1000				85.07	72.26	51.56	86.55	75.54	54.50
2000				86.96	77.15	58.49	87.92	79.85	61.62
(<) 5000				88.32	81.48	65.54	89.14	84.28	70.33

Tab. 2. Lexical Coverage for Czech

3.4 Processing the Czech Data

The proportion of the dictionary headword list (in %) covered by the respective subcorpus has been observed. All results are displayed in Table 2.

For each sampled subcorpus, three values are presented, representing the subcorpus lexical coverage of the dictionary headword list if at least one, ten, and one hundred occurrences of lexical items in the corpus are required, respectively. For example, the 100 M subcorpus sampled from syn 20xx covers 79.65% of the dictionary headword list on condition that 1 corpus occurrence is considered satisfactory, but only 31.44% if at least 100 corpus occurrences are required.

The values from the table are visualized in Fig. 1. The *x* axis represents the corpus size in millions of tokens and the *y* axis shows the coverage of vocabulary (in %) by the respective (sub)corpora. As the left part of the graph is rather dense, the situation with smaller subcorpora is better visible if corpus size is plotted in a logarithmic scale (Fig. 2).

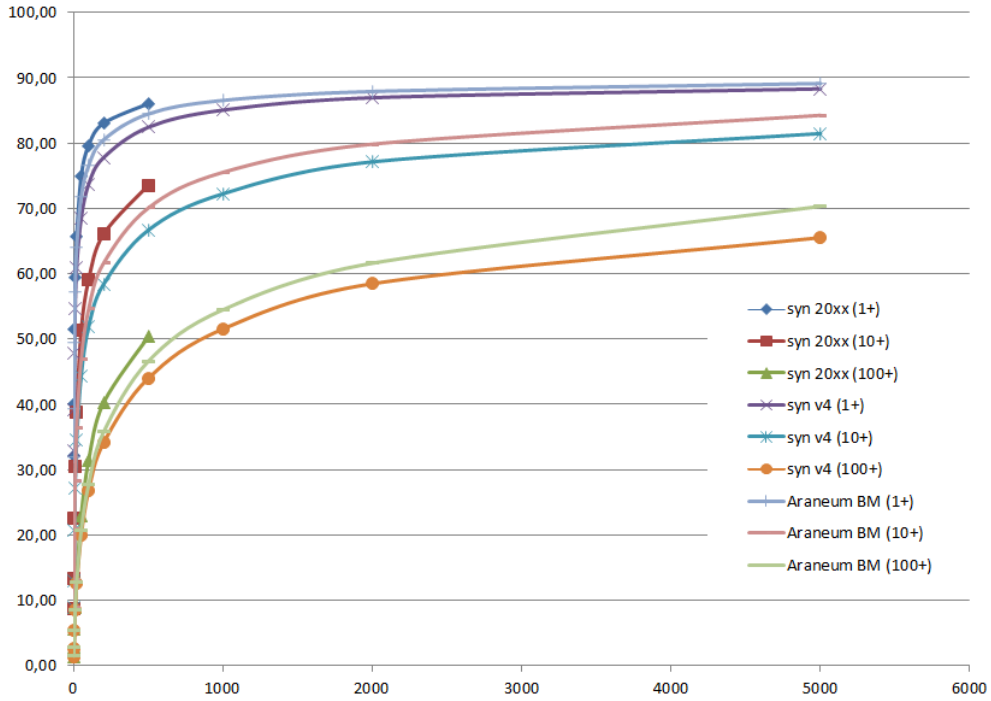

Fig. 1

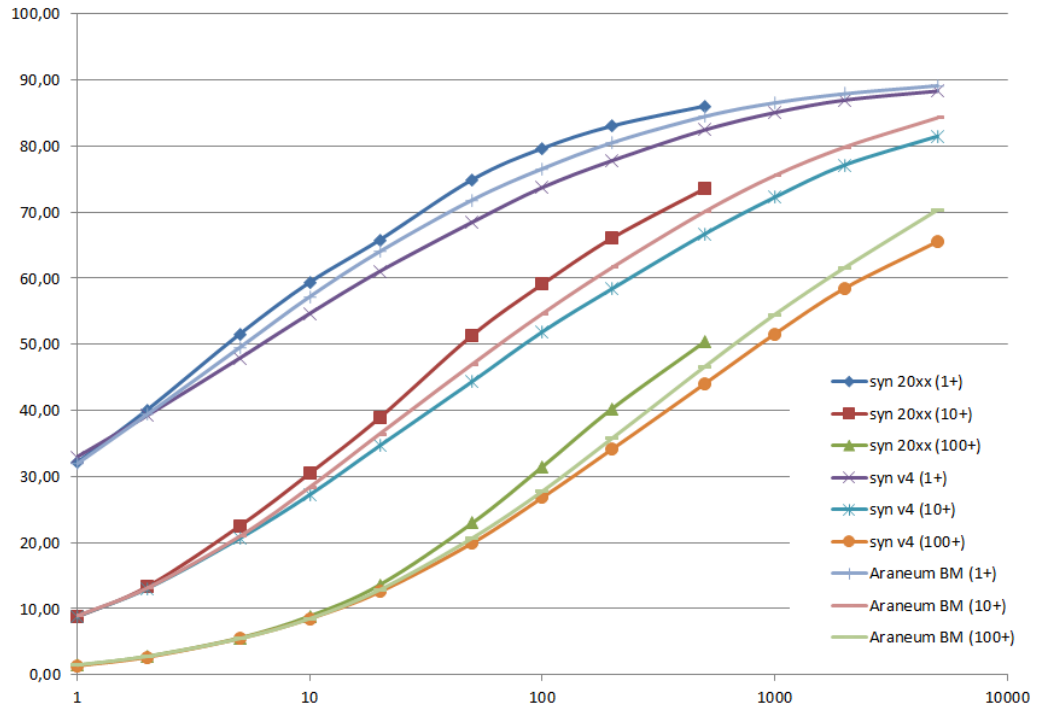

Fig. 2

3.5 The Slovak Data

The procedure for Slovak was similar to that of Czech, with the main difference being the sizes of both traditional and web corpora. The respective results are summarized in Table 3.

Corpus size (M)	pri,m-6.1-vyv			prim-6.1-all			Araneum SM		
	(1+)	(10+)	(100+)	(1+)	(10+)	(100+)	(1+)	(10+)	(100+)
1	31.66	8.38	1.26	31.47	8.59	1.30	30.50	8.69	1.41
2	39.30	13.06	2.50	38.08	12.57	2.54	37.84	12.93	2.72
5	52.00	22.43	5.26	49.16	20.58	5.14	48.62	20.60	5.29
10	59.66	30.32	8.62	56.58	28.03	8.18	55.74	27.59	8.33
20	66.69	39.13	13.51	64.67	36.64	12.54	62.10	35.61	12.57
50	74.64	51.63	22.85	73.05	49.08	21.22	69.55	45.77	20.11
100	78.62	59.97	31.25	77.28	56.84	29.00	74.31	53.19	27.09
200	81.58	66.88	40.23	80.76	64.30	37.45	77.99	59.90	35.00
(<) 500				83.68	72.07	48.94	81.76	68.04	45.37
(<) 1000				84.83	75.82	55.36	83.64	72.89	52.94
2000							84.98	76.02	58.12
3500							85.22	77.84	59.76

Tab. 3. Lexical Coverage for Slovak

The figures show similar progress as those for Czech, forming the shapes displayed at Fig. 3 (in linear scale for subcorpora sizes), and Fig. 4 (logarithmic scale).

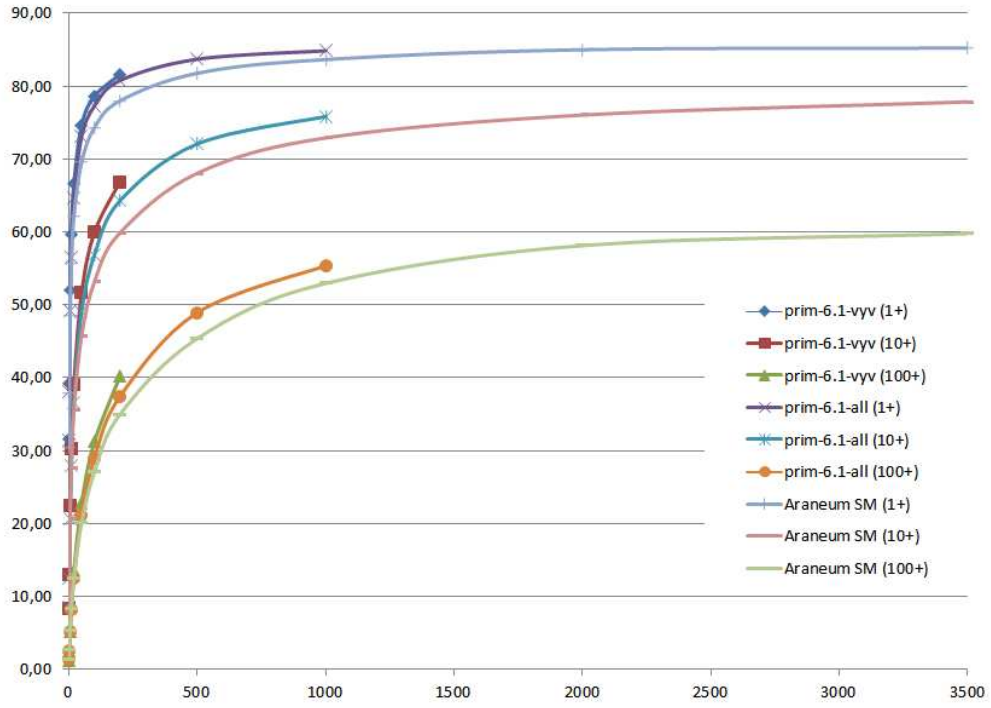

Fig. 3

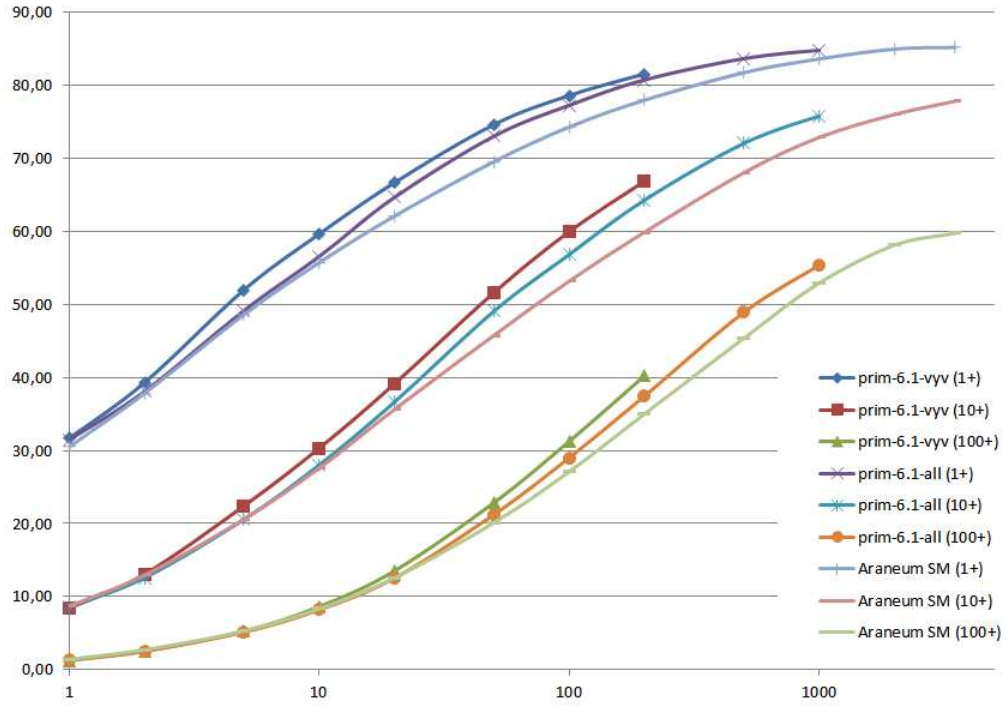

Fig. 4

4 Conclusion and Further Work

The results are mostly consistent with our expectations, and can be summarized as follows:

(1) The lexical coverage for both languages is growing steeply with the size of corpus for smaller corpora, but a saturation can observed at approximately 1 billion tokens.

(2) The coverage of the Czech headword list approaches 90%, while the Slovak one stops at approximately 85%, which deserves a more detailed analysis. The quick lookup reveals several cases here: the Czech headword lists contained many regular derivates from infrequent words, spelling variants not present in contemporary language, and even typos in the retro-digitized dictionary); the unmatched items in the Slovak list also contain a large number in geographical and inhabitant names that rarely occur in text.

(3) Both balanced corpora are slightly "better" within the range of their size, this advantage can be outperformed be the sheer size of larger corpora.

(4) Traditional unbalanced corpus is slightly "worse" in smaller sizes for Czech and slightly "better" for Slovak. The difference, however, almost disappears with corpora larger than 2 billion tokens.

(5) As a source for lexicographic work, (at least) 2 Gigaword corpus is to be recommended.

More research is necessary to evaluate the differences between traditional and web corpora, most notably in text types, domains, genres and registers, as well as with wordlist derived from different dictionaries.

Acknowledgement

The research described in this paper has been supported by VEGA Grant Agency, Project No. 2/0017/17.

References

Baroni M., Bernardini, S., Ferraresi A., Zanchetta E. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. Language Resources and Evaluation 43 (3), pp. 209–226.

Benko, V. 2014. Aranea: Yet another Family of (Comparable) u Corpora. In Text, Speech, and Dialogue. 17th International Conference, TSD 2014 Brno, Czech Republic, September 8 – 12, 2014, Proceedings. Ed. P. Sojka et al. – Cham – Heidelberg – New York – Dordrecht – London : Springer, 2014, 21–29. ISBN 978-3-319-10816-2.

Jakubíček, M. – Kilgarriff, A. – Kovář, V. – Rychlý, P. – Suchomel V. 2013. The TenTen Corpus Family. In 7th International Corpus Linguistics Conference, Lancaster, July 2013.

Horák, G. et al. (Ed.) 1981, Česko-slovenský slovník. Bratislava : Veda, 1981.

Khokhlova, M. 2016. Large Corpora and Frequency Nouns. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016", Moscow, June 1–4, 2016.

Kilgarriff, A. 2001. Comparing Corpora. International Journal of Corpus Linguistics, 6(1), 97–133.

Kilgarriff, A. et al. 2004. The Sketch Engine. In: G. Williams and S.Vessier (eds.), Proceedings of the eleventh EURALEX International Congress EURALEX 2004 Lorient, France, July 6-10, 2004. Lorient : Université de Bretagne-Sud, pp. 105–116.

Kilgarriff, A., Rychlý, P., Jakubíček, M., Kovář, V., Baisa, V., Kocincová, L. 2014. Extrinsic Corpus Evaluation with a Collocation Dictionary Task. Proc. LREC 2014. Reykjavik, pp. 545-552.

Křen, M., Cvrček, V., Čapka, T. et al., 2016, SYN v4: large corpus of written Czech, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, http://hdl.handle.net/11234/1-1846.

PSP 2000. Považaj, M. (Ed.): Pravidlá slovenského pravopisu. 3., upravené a doplnené vydanie. Bratislava : Veda 2000.

Schäfer, R. – Bildhauer, F. 2013. Web Corpus Construction. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Šimková, M. – Garabík, R.: Slovenský národný korpus (2002–2012): východiská, ciele a výsledky pre výskum a prax. In: Jazykovedné štúdie XXXI. Rozvoj jazykových technológií a zdrojov na Slovensku a vo svete (10 rokov Slovenského národného korpusu). Ed. Katarína Gajdošová — Adriána Žáková. Bratislava: VEDA 2014, 35– 64.

SNK 2013a. Slovenský národný korpus – prim-6.1-public-all. Bratislava: Jazykovedný ústav Ľ. Štúra SAV, http://korpus.juls.savba.sk.

SNK 2013b. Slovenský národný korpus – prim-6.1-public-vyv. Bratislava: Jazykovedný ústav Ľ. Štúra SAV, http://korpus.juls.savba.sk.

Web Corpora – the best possible solution for tracking rare phenomena in underresourced languages: clitics in Bosnian, Croatian and Serbian

Edyta Jurkiewicz-Rohrbacher
University of Helsinki,
Universität Regensburg
edyta.jurkiewicz@helsinki.fi

Zrinka Kolaković
Universität Regensburg
zrinka.kolakovic@ur.de

Björn Hansen
Universität Regensburg
bjoern.hansen@ur.de

Abstract

Complex linguistic phenomena, such as Clitic Climbing in Bosnian, Croatian and Serbian, are often described intuitively, only from the perspective of the main tendency. In this paper, we argue that web corpora currently offer the best source of empirical material for studying Clitic Climbing in BCS. They thus allow the most accurate description of this phenomenon, as less frequent constructions can be tracked only in big, well-annotated data sources. We compare the properties of web corpora for BCS with traditional sources and give examples of studies on CC based on web corpora. Furthermore, we discuss problems related to web corpora and suggest some improvements for the future.

1 Introduction

One of the main goals of modern electronic text corpora is providing linguists with tools that would allow them to verify their theories or hypotheses, and eventually to make new findings on language in a quick and efficient way, without having to use intuition-based research methods, which are prone to bias. We share the view of Gries and Newman (2013, 253) that "over the last few decades, corpus-linguistics methods have established themselves as among the most powerful and versatile tools to study language acquisition, processing, variation and change". In the theoretical literature, grammaticality of constructions is often assessed according to the scholar's intuition. Less-frequent phenomena are often only vaguely glimpsed, or in most cases evaluated as incorrect.

In the present paper, we show how web corpora can help settle disputes concerning such rare phenomena, lead to solid discoveries, and correct often inconsistent theoretical claims. As our point of departure we take contradictory theoretical claims related to clitics (CLs) in Bosnian, Croatian and Serbian (BCS), which partially arise from the lack of solid empirical data in research. As examples of this, we consider the case of pronominal and reflexive CCs in BCS which climb out of complement clauses into higher clauses: a phenomenon called Clitic Climbing (CC). Web corpora – linguistically annotated and available via on-line corpus managers – appear to be a very convenient source of data, in particular for those studying underresourced languages like BCS[1].

Here, we argue that for the purposes of studying the constraints on CC out of *da*-complements and multiply embedded infinitive complements in BCS, the corpora compiled from top domains {bs,hr,sr}WaC (Ljubešić and Klubička, 2014) are currently a better source of authentic data for BCS when it comes to size, available metainformation and searchability than traditionally compiled sources.

Finally, we comment on problems that linguists face while working with web corpora. Moreover, we present some suggestions for corpus designers that, in our view, could improve the reliability of linguistic studies and the precision of queries.

2 Clitic climbing in BCS

One possible definition of CLITIC CLIMBING concerns "a construction in which the clitic is associated with a verb complex in a subordinate clause but is actually pronounced in constructions with a higher predicate" (Spencer and Luìs, 2012, 162). The classical example of CC out of a *da*-complement is given below, where the CL *ih* 'them' generated by the *da*-complement *čita* 'reads' appears in the second position in the sentence (the so-called Wackernagel position):

[1] As recognized by the group of linguists behind the Regional Linguistic Data Initiative; for more information see https://reldi.spur.uzh.ch

(1) *Niko* ***ih***$_2$ *ne* *može*$_1$ *da*
Nobody them.ACC NEG can.3PRS COMP
čita$_2$.
read.PRS
'Nobody can read them.' (Marković, 1955, 38)[2]

Nevertheless, CC is not always realized in BCS, as we observe in the empirical material. (2) and (3) provide examples of the Serbian semifinite *da*-complements, consisting of the complementizer-like element *da* and a verbal form coinciding with the present tense form, which is the counterpart of the infinitive complement. In both cases, the complement-embedding predicate *sm(j)eti* 'to be allowed' is the matrix verb and *dozvoliti* 'to allow' is a part of the *da*-complement. In each sentence, the pronominal CL *im* 'them' appears as the complement of the semifinite verb. In contrast to (2), where the CL stays in the clause together with its governor, in (3) the CL climbs out of the embedded *da*-complement in which it was generated into the clause with the higher predicate.

(2) *Ne bismo smeli*$_1$ *da*
NEG cond.1PL be.allowed.PTCP.PL.M COMP
im$_2$ *dozvolimo*$_2$ *(...)*
them.DAT allow.1PRS
'We should not allow them (to do) that (...).' (srWaC v1.2)

(3) *To* ***im***$_2$ *Vučić ne*
that.ACC them.DAT Vučić NEG
sme$_1$ *da* *dozvoli*$_2$.
be.allowed.3PRS COMP allow.3PRS
'Vučić must not allow them (to do) that' (srWaC v1.2)

The second context in which we observe different positions of CLs is multiply embedded infinitive complements. While in (4) the CL *mi* 'me' generated by *uskratiti* 'to deprive' stays in situ, in (5) the CC *ga* 'him' climbs out of its infinitive complement *dati* 'give' over the infinitive complement *odbiti* 'refuse' and takes second position within the matrix clause.

(4) *(...) možete* ***si***$_1$ *dozvoliti*$_2$ *uskratiti*$_3$
can.2PRS REFL.DAT allow.INF deprive.INF
mi$_3$ *sve*
me.DAT everything
'(...) you can allow yourselves to deprive me of everything (...)' (hrWaC v2.2)

(5) *(...) a ti* ***ga***$_3$ *imaš*$_1$ *pravo*$_1$
and you it.ACC have.2SG right.ACC
odbiti$_2$ *dati*$_3$.
refuse.INF give.INF
'(...) and you have the right to refuse to give it.' (hrWaC v2.2)

As we shall see in the next section, the latter phenomenon has been studied only by Hansen et al. (In press), while the former is discussed only in a few studies or vaguely mentioned in studies dedicated to other phenomena related to CLs. All in all, information found in literature is based mainly on a few, mostly self-produced examples and, as we will show in the next section, the conclusions drawn by different scholars are highly contradictory.

3 Related work

Some authors argue that CC out of *da*-complements is strictly impossible (Ćavar and Wilder, 1994; Browne, 2003, 41), emphasizing that CLs in *da*-complements have to directly follow *da* and precede the semifinite verb (see Browne 2003: 41). Others, however, do accept it, albeit with some additional remarks. Stjepanović (Stjepanović, 2004, 174ff) argues that *da*-complements allow CC in a similar way to infinitival clauses, but while discussing examples with CLs that have climbed out of *da*-complements, she rather vaguely admits that these "are acceptable sentences, however, they are short of perfect" (Stjepanović, 2004, 201). A similar perspective is presented by Franks and King (2000, 253). Bošković (2001, 3) claims that "South Slavic systems also involve clitic climbing operations out of finite clauses", but all his examples which should support that claim are marked with a question mark. Finally, Progovac (2005, 146) admits that "some speakers of Serbian" do not accept her data, i.e. do not accept CC in these contexts.

In contrast to above mentioned authors, Marković (Marković, 1955) analysed CC

[2]The matrix is always indexed with 1, while complement predicates are indexed with 2, (if there are more, then also with 3 etc.). CLs are indexed according to their governors so that their climbing can be traced. Additionally, CLs are marked with bold.

of pronominal and reflexive CLs out of *da*-complements in naturally occurring sentences. In his opinion, the variation in clitic positioning is closely related to the (at the time) recent and increased tendency to suppress the infinitive as a complement by replacing it with a *da*-complement (Marković, 1955, 40). Furthermore, he claimed that ekavian Serbs preferred to keep the pronominal CL directly after *da* instead of moving it as close as possible to the second position in the sentence (Marković, 1955, 39). Still, he emphasized a certain degree of variation in the middle and western language area of Serbia, where cases of CLs placed left of *da* were attested (Marković, 1955, 37). Besides this diatopic variation factor, he noted that diaphasic variation plays a role as well, since pronominal CLs preceding *da* may often be found in journalistic texts published in Sarajevo and in Serbian *belles lettres* (Marković, 1955, 35).

As CC has been studied in more detail for Czech than for BCS, we looked into the findings concerning this Slavonic language. Many scholars who have written on CC in Czech have noticed consistent patterns linked to different types of matrix verbs. They have observed that in the case of infinitive complements Czech pronominal and reflexive CLs can climb out of infinitives which are governed by raising and subject control matrix verbs, while some additional restrictions occur in the case of object control[3] (George and Toman, 1976; Dotlačil, 2004; Rezac, 2005; Hana, 2007). Furthermore, while above mentioned authors argue that in certain cases CC out of object-controlled infinitives is possible, others completely reject such a possibility (Thorpe, 1991; Junghanns, 2002). It is important to note once more that even in the case of studies of CC in Czech the majority of scholars based their statements on self-constructed examples. As far as we know, no serious corpus study with inferential statistical methods has been undertaken yet.

While there are many studies on CC out of infinitive complements, especially for Czech, and many theories about constraints which prevent CLs from climbing into higher clauses have been postulated, there has been only one study in which the position of CLs in the context of multiply embedded infinitive complements was examined and compared in BCS (Hansen et al., In press).

We believe that corpora are the perfect environment for verifying the above mentioned theoretical claims and for forming hypotheses on understudied phenomena. This is because they contain sentences in their natural environment, so the possibility of bias in evaluation of correctness is minimal in comparison to the informal acceptability judgements of authors or to questionnaire-based methods.

Furthermore, since in corpora sentences occur in their natural context and are not adjusted to the context of interest, the ecological validity (degree of similarity between the study and the authentic context) of the results is higher than in laboratory environments. We thus assume that an ideal triangulation of methods should combine corpus with additional experimental data in order to avoid the problem of negative evidence. Our first goal is to test whether the relation between the matrix verb and the position of CCs generated in the embedded *da*-complements is statistically significant and whether any tendencies regarding CC out of stacked infinitive complements can be detected.

4 Corpora of BCS – an overview

Among the three languages in focus, construction of a national corpus has so far begun only for Croatian (Croatian National Corpus HNK, (Tadić, 2009)). The biggest traditionally compiled corpus of Serbian is the Corpus of Contemporary Serbian Language (SrpKor2013) developed at the Faculty of Mathematics of the University of Belgrade by Miloš Utvić and Duško Vitas. In a sense, SrpKor2013 has taken on the role of the national corpus. As of today, no national corpus of Bosnian has been built. The only traditional, monolingual source is the Oslo Corpus of Bosnian Text (OCTB) (Santos, 1998).

The main features of the most relevant sources of contemporary texts written originally in BCS are summarized below.

From Table 1 it may be seen that most corpora can be queried through Corpus Query Processor-based engines or similar, but in most cases access to meta-information is very limited. Only HNK

[3]The raising-control dichotomy is represented in the following way: "i) semantically, raising verbs have one argument fewer than the corresponding control verbs, e.g., *seem* is a (semantically) 1-argument verb, while *try* is a (semantically) 2-argument verb; ii) structurally, the raised argument and the subject of the infinitival verb are the same element [...], while the controller and the subject of the infinitival verb are two different elements"(Przepiórkowski and Rosen, 2005).

	size (tokens)	lemmatized	POS	MSD	text type	Query type
Bosnian						
OCBT	1,500,000	yes	no	no	fiction, essays, newspapers, children's books, Islamic texts, legal texts, folklore	CQP
Croatian						
HNK	2,559,160	yes	yes	yes	Croatian literature: novels, stories, essays, diaries, (auto)biographies non-fiction: newspapers, magazines, journals, brochures, correspondence	CQL
Hrvatska jezična riznica developed at the Institute for Croatian Language and Linguistics in Zagreb	no data	no	no	no	Croatian literature, non-fiction: scientific publications, online journals and newspapers	
Serbian						
InterCorp v9 - Serbian (Latin) (subocorpus of original Serbian texts)	563,782	yes	yes	yes	literature	CQL
SrpKor2013	122,255,064	yes	yes	no	administrative, journalism, literature, academic, other	CQP

Table 1: The most important traditionally compiled corpora of BCS.

and InterCorp have been morphosyntactically annotated.

The three web corpora for BCS, on the other hand, are quite impressive when it comes to size, searchability and meta-information, as summarized in Table 2 on the next page.

The annotation process has not been revised but its estimated accuracy is quite promising as it reaches the level of 92.33%-92.53% as regards morphosyntactic tagger performance and 97.86%-98.11% as regards part-of-speech tagger accuracy (Ljubešić et al., 2016, 4268).

Generally, the main objection against web corpora as a source of data for linguistic studies, in comparison to traditionally compiled sources, is held to be the lack of control of text variety and the high level of author anonymity. While the former issue can be partially solved by specifying particular domains or by direct reference to the source web page, the latter issue seems currently unsolvable. Even consultation of a source web page does not guarantee correct identification of an author's social background, in particular their native language, place of origin or age. The linguist should bear in mind that some caution is needed with respect to linguistic variation.

The problem of control concerns not only web corpora, but any kind of big data. Although SrpKor2013 and HNK theoretically allow for the control of functional style, they lack a proper specification which would include a description of the actual balance between different text types. Therefore, in respect of text variety control, large traditionally compiled corpora turn out to be as similarly imperfect a source as {bs,hr,sr}WaC.

On the other hand, we are aware that some trials of automatic genre analysis have been carried out and are summarized in Mehler et al. (2010). Among Slavonic languages, the most recent solution has been proposed in the Czech National Corpus by Cvrček (2017), who following Biber (1991) and Biber and Conrad (2009) employed multidimensional analysis of text varieties in the 9,000,000-word corpus.

5 CC and web corpora

As mentioned in the Introduction, in the case of pronominal and reflexive CLs certain positions of CLs seem to be preferred in particular constructions. As a consequence, scholars may consider the less-frequent position to be unacceptable. Corpora can help determine the circumstances under which the rarely occurring CL position can be realized as long as a sufficient number of accurate examples can be retrieved.

The crucial factor here is size. For example, a search of CC out of *da*-complements in Serbian yields only two examples in the literary part of InterCorp v9. srWac uses the same tagset, so a comparable query can be conducted. However, due to its enormous size, the search must be performed separately for each matrix verb. The results of a study conducted on 15 verbs belonging to three different syntactic types enable us to form the hypothesis that CC is marginally possible with raising and subject control types of matrix verbs (the Chi-square test for independence between syntactic type and CC yields a significant p-value = 7.948e-11) and its frequency related to overall frequency of *da*-complements varies between 0.0116 and 0.0009.

In the case of multiply embedded infinitive complements, it turned out that reflexivity of the infinitive that embeds further infinitives plays a crucial role in preventing CC (an Odds Ratio test with a 95% confidence level yields 502.8000, p<0.0001). This conclusion could not be made on the basis of traditional sources as either they are too small or the rare constructions could not be retrieved due to lack of meta-information.

As the three web corpora use the same tagset, the very same searches can be conveniently applied to all three languages and the variation in the distribution of constructions with and without CC can be easily examined across languages. This, for example, allowed Hansen et al. (In press) to find that CC out of complements containing stacked infinitives is similarly distributed in all three languages.

For both constructions, web corpora also allowed the formulation of hypotheses that can be further examined in assessment tests. For example, with respect to the reflexivity constraint detected in the study of stacked infinitive complements, we can test whether different types of reflexives (lexical, reciprocal, reflexive occupying the place of direct/indirect object) are equally important in blocking CC. In the case of CC out of *da*-complements the acceptability of CC in the context of raising and subject controlled predicates can be tested with respect to diatopic variation (since those data are missing from Web Corpora) in order

	size (tokens)	lemmatized	POS	MSD	Query type
bsWaC v1.2	248,478,730	yes	yes	yes	CQL
hrWaC v1.2	1,210,021,198	yes	yes	yes	CQL
srWaC v1.2	554,627,647	yes	yes	yes	CQL

Table 2: BCS corpora compiled from .bs, .hr and .sr top level domains.

to prove Marković's (1955) claims.

6 Suggestions for improvements in corpus design

As shown above, web corpora are currently the most promising source of data for studying the competing positions of CLs in BCS. They provide empirical evidence for claims often rejected in the literature on the subject.

The necessary condition for such a study is satisfactory corpus size. However, this condition is not sufficient without appropriate tools for searching through big data. The handful of traditionally compiled corpora for BCS do not, in most cases, fulfil the first condition, or they do not provide enough meta-information to allow accurate searches to be conducted.

On the other hand, currently available web corpora satisfy the size condition. The unified tagset and search mechanism allow comparable queries to be conducted in all three languages.

The two main problems concerning web corpora are control for text-types and the question of reliability of obtained results. We are aware that neither of those problems can be solved easily. From the linguistic point of view, we suggest that more attention should be paid to developing methods that would allow texts to be classified by functional style as mentioned in Section 4.

Also the evaluation of search reliability leaves plenty of room for improvement as currently no gold standards are available. While the precision of queries can be evaluated by means of extrapolations based on samples as suggested by Sean Wallis[4] , no recommendations have been offered so far about the assessment of recall.

Of course, the quality of results depends on the complexity and the accuracy of annotation. The ambiguity of queries could be decreased through tagging of syntactic features or through sentence clause identification, which, in the case of English, has recently been under development by Muszyńska (2016) and Niklaus et al. (2016) but seems to still be an undeveloped topic as regards Slavonic languages.

[4] https://corplingstats.wordpress.com/2014/04/10/imperfect-data/

Acknowledgments

This work was financed by DFG 'Microvariation of the Pronominal and Auxiliary Clitics in Bosnian, Croatian and Serbian. Empirical Studies of Spoken Languages, Dialects and Heritage Languages' (HA 2659/6-1, 2015-2018). The authors gratefully thank to the Reviewers for the comments and recommendations which helped to improve the readability and quality of the paper.

References

Douglas Biber and Susan Conrad. 2009. *Register, Genre and Style*. Cambridge University Press, New York, NY.

Douglas Biber. 1991. *Variation across speech and writing*. Cambridge University Press, Cambridge.

Željko Bošković. 2001. *On the nature of the syntax-phonology interface: cliticization and related phenomena*. Elsevier, Amsterdam.

Wayles Browne. 2003. Razlike u redu riječi u zavisnoj rečenici. *Wiener Slawistischer Almanach*, 57:39–44.

Damir Ćavar and Chris Wilder. 1994. "Clitic third" in Croatian. *Linguistics in Potsdam*, 1:25–63.

Niklaus Christina, Bernhard Bermeitinger, Siegfried Handschuh, and Andr´e Freitas. 2016. A sentence simplification system for improving relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 170–174, Osaka. COLING.

Václav Cvrček, Zuzana Komrsková, David Lukeš, Petra Poukarová, Anna Řehořková, and Adrian Zasinas. 2017. Genre variation in interactions. Paper presented at Interakce v socio-kognitivní, antropologické a historické perspektivě, Prague.

Jakub Dotlačil. 2004. The syntax of infinitives in Czech. Master's thesis, University in Tromsø, Tromsø.

Steven Franks and Tracy H. King. 2000. *A Handbook of Slavic clitics*. Oxford University Press, Oxford.

Leland George and Jindrich Toman. 1976. Czech clitics in universal grammar. In Salkiko S. Mufwene, Carol A. Walker, and Sanford B. Steever, editors, *Papers from the 12th Regional Meeting Chicago Linguistic Society*, pages 235–249. Chicago Linguistic Society, Chicago.

Stefan Th Gries and Newman. 2013. Creating and using corpora. In Robert J. Podesva and Devyani Sharma, editors, *Research Methods in Linguistics*, pages 257–287. Cambridge University Press, Cambridge.

Jirka Hana. 2007. *Czech Clitics in Higher Order Grammar*. Ph.D. thesis, The Ohio State University, Ohio.

Björn Hansen, Zrinka Kolaković, and Edyta Jurkiewicz-Rohrbacher. In press. Clitic climbing and infinitive clusters in Bosnian, Croatian and Serbian – a corpus-driven study. In Eric Fuß, Marek Konopka, Beata Trawiński, and Ulrich H. Waßner, editors, *Grammar and Corpora 2016*. Heidelberg University Publishing (heiUP), Heidelberg.

Uwe Junghanns. 2002. Clitic climbing im Tschechischen. *Linguistische Arbeitsberichte*, 80:57–90.

Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg.

Nikola Ljubešić, Filip Klubička, Željko Agić, and Iwo-Pavao Jazbec. 2016. New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4264–4270, Paris. ELRA.

Svetozar Marković. 1955. Položaj zamjeničke enklitike u vezi sa naporednom upotrebom infinitiva i prezenta sa svezicom da. *Naš Jezik*, 6(1–2):33–40.

Alexander Mehler, Serge Sharoff, and Marina Santini, editors. 2010. *Genres on the Web: Computational Models and Empirical Studies*. Springer, Dordrecht.

Ewa Muszyńska. 2016. Graph- and surface-level sentence chunking. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics - Student Research Workshop*, pages 93–99, Berlin. ACL.

Liljana Progovac. 2005. *A Syntax of Serbian: Clausal Architecture*. Slavica Publishers, Bloomington.

Adam Przepiórkowski and Alexandr Rosen. 2005. Czech and Polish raising/control with or without structure sharing. *Research in Language*, 3:33–66.

Milan Rezac. 2005. The syntax of clitic climbing in Czech. In Lorie Heggie and Francisco Ordóñez, editors, *Clitics and affix combinations. Theoretical perspectives*, pages 103–140. Benjamins, Amsterdam.

Diana Santos. 1998. Providing access to language resources through the World Wide Web: the Oslo Corpus of Bosnian Texts. In *Proceedings of The First International Conference on Language Resources and Evaluation*, pages 475–481.

Andrew Spencer and Ana R. Luìs. 2012. *Clitics: An Introduction.* Cambridge University Press, Cambridge.

Sandra Stjepanović. 2004. Clitic climbing and restructuring with "finite clause" and infinitive complements. *Journal of Slavic Linguistics*, 12(1):173–212.

Marko Tadić. 2009. New version of the Croatian National Corpus. In Dana Hlaváčková, Aleš Horák, Klára Osolsobě, and Pavel Rychlý, editors, *After Half a Century of Slavonic Natural Language Processing*, pages 199–205. Masaryk University, Brno.

Alena Irena Thorpe. 1991. *Clitic placement in complex sentences in Czech*. Ph.D. thesis, Brown University, Rhode Island.

Removing Spam from Web Corpora Through Supervised Learning Using FastText

Vít Suchomel
Natural Language Processing Centre
Faculty of Informatics, Masaryk University, Brno, Czech Republic
xsuchom2@fi.muni.cz

Abstract

Unlike traditional text corpora collected from trustworthy sources, the content of web based corpora has to be filtered. This study briefly discusses the impact of web spam on corpus usability and emphasizes the importance of removing computer generated text from web corpora.

The paper also presents a keyword comparison of an unfiltered corpus with the same collection of texts cleaned by a supervised classifier trained using FastText. The classifier was able to recognise 71 % of web spam documents similar to the training set but lacked both precision and recall when applied to short texts from another data set.

1 Web Spam in Text Corpora

It has been shown that boilerplate, duplicates, and spam skew corpus based analyses and therefore have to be removed, see nonsense examples of word use in an application for English learners based on a web corpus in figure 1. While the first two issues have been successfully addressed, e.g. by (Marek et al., 2007; Pomikálek, 2011; Versley and Panchenko, 2012; Schäfer and Bildhauer, 2013), spam might be still observed in web corpora as reported by (Kilgarriff and Suchomel, 2013). It was spam that represented the main difference between their 2008 and 2012 corpora crawled from the web. That is why a spam cleaning stage should be a part of the process of building web corpora.

The traditional definition of web spam is *actions intended to mislead search engines into ranking some pages higher than they deserve* (Gyöngyi and Garcia-Molina, 2005). The Google document 'Fighting Spam'[1] describes the kinds of spam that Google finds, and what they do about it.

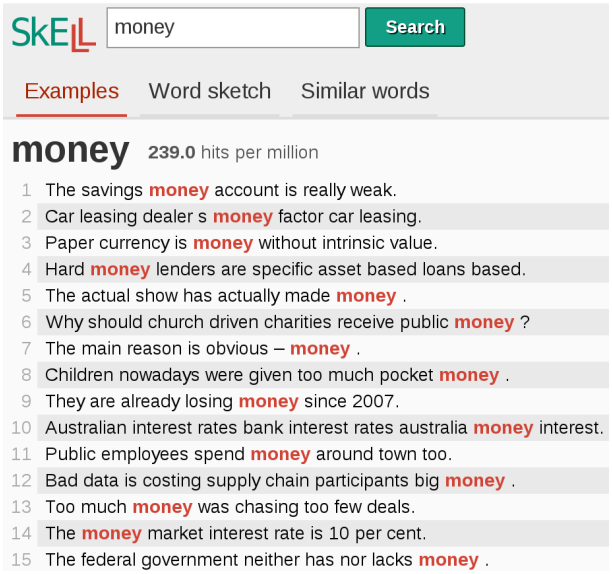

Figure 1: Web spam in examples of use of word 'money' at skell.sketchengine.co.uk – see lines 2, 4 and 10.

Text alteration techniques consist in changing the frequency properties of a web page content in favour of spam targeted words or phrases: *repetition of terms related to the spam campaign target, inserting a large number of unrelated terms, often even entire dictionaries, weaving of spam terms into contents copied from informative sites, e.g. news articles, glueing together sentences or phrases from different sources* as reported by (Gyöngyi and Garcia-Molina, 2005).

Automatically generated content does not provide examples of authentic use of a natural language. Nonsense, incoherent or any unnatural texts such as the following short instance have to be removed from a good quality web corpus: *Edmonton Oilers rallied towards get over the Montreal Canadiens 4-3 upon Thursday.Ryan Nugent-Hopkins completed with 2 aims, together with*

[1] https://www.google.com/insidesearch/howsearchworks/fighting-spam.html

the match-tying rating with 25 seconds remaining within just legislation.[2]

The following types of automatically generated content are examples of documents penalised by Google:[3] *Text translated by an automated tool without human review or curation before publishing. Text generated through automated processes, such as Markov chains. Text generated using automated synonymizing or obfuscation techniques.* These kinds of spam should certainly be eliminated from web corpora while the other two examples given by Google may not present a harm to the corpus use: *Text generated from scraping Atom/RSS feeds or search results. Stitching or combining content from different web pages without adding sufficient value.*

In contrast to the traditional or search engine definitions of web spam, the corpus use point of view is not concerned with intentions of spam producers or the justification of the search engine optimisation of a web page. A text corpus built for NLP or linguistics purpose should contain coherent and consistent, meaningful, natural and authentic sentences in the target language. Only texts created by spamming techniques breaking those properties should be detected and avoided. The unwanted non-text is this: computer generated text, machine translated text, text altered by keyword stuffing or phrase stitching, text altered by replacing words with synonyms using a thesaurus, summaries automatically generated from databases (e.g. stock market reports, weather forecast, sport results – all of the same kind very similar), and finally any incoherent text. Varieties of spam removable by existing tools, e.g. duplicate content, link farms (quite a lot of links with scarce text), are only a minor problem.

Avoiding web spam by selecting trustworthy corpus sources such as Wikipedia, news sites, government and academic webs works well: (Baisa and Suchomel, 2014) show it is possible to construct medium sized corpora from URL whitelists and web catalogues. (Spoustová and Spousta, 2012) used a similar way of building a Czech web corpus. Also the BootCaT method (Baroni and Bernardini, 2004) indirectly avoids spam by relying on a search engine to find non-spam data. Despite the avoiding methods being successful, it is doubtful a huge web collection can be obtained just from trustworthy sources.

Furthermore, language independent methods of combating spam might be of use. (Ntoulas et al., 2006) reported web spamming was not only a matter of the English part of internet. Spam was found in their French, German, Japanese and Chinese documents as well.

2 Removing Spam Using a Supervised Classifier

This section describes training and evaluation of a supervised classifier to detect spam in web corpora.

We have manually annotated a collection of 1630 web pages from various web sources from years 2006 to 2015.[4] To cover the main topics of spam texts observed in our previously built corpora, we included 107 spam pages promoting medication, financial services, commercial essay writing and other subjects. Both phrase level and sentence level incoherent texts (mostly keyword insertions, n-grams of words stitched together or seemingly authentic sentences not conveying any connecting message) were represented. Another 39 spam documents coming from random web documents identified by annotators were included. There were 146 positive instances of spam documents altogether.

The classifier was trained using FastText (Joulin et al., 2016) and applied to a large English web corpus from 2015. The expected performance of the classifier was evaluated using a 30-fold cross-validation on the web page collection. Since our aim was to remove as much spam from the corpus as possible, regardless false positives, the classifier confidence threshold was set to prioritize recall over precision. The achieved precision and recall were 71.5 % and 70.5 % respectively. Applying this classifier to an English web corpus from 2015 resulted in removing 35 % of corpus documents still leaving enough data for the corpus use.

An inspection of the cleaned corpus revealed the relative count of usual spam related keywords dropped significantly as expected while general words not necessarily associated with spam were affected less as can be seen in table 1.

Another evaluation of the classifier was performed by manually checking 299 random web documents from the cleaned corpus and 25 ran-

[2] `http://masterclasspolska.pl/forum/`

[3] Google quality guidelines – `https://support.google.com/webmasters/answer/2721306`

[4] The collection is a part of another classification experiment by the same authors not covered by this paper.

dom spam documents removed by the classifier. The achieved precision was 40.0 % with the recall of 27.8 %. The error analysis showed the classifier was not able to recognise non-text rather than spam. 17 of 26 unrecognised documents were scientific paper references or lists of names, dates and places, i.e. *Submitted by Diana on 2013-09-25 and updated by Diana on Wed, 2013-09-25 08:32* or *January 13, 2014 January 16, 2014 Gaithersburg, Maryland, USA*. Such web pages were not present in the training data since we believed it had been removed from the corpus sources by a boilerplate removal tool and paid attention to longer documents. Not counting these 17 non-text false negatives, the recall would reach 52.6 %.

To find out what was removed from the corpus, relative counts of lemmas[5] in the corpus were compared with the BNC[6] in figures 2 and 3. A list of lemmas in the web corpus with the most reduced relative lemma count caused by removing unwanted documents is presented in 4.

The inspection showed there were a lot of spam related words in the original web corpus and that spam words are no longer characteristic of the cleaned version of the corpus in comparison to the BNC.[7]

3 Conclusion

We view computer generated text as the main kind of spam decreasing the quality of web corpora. A classifier trained on spam documents was applied to remove unwanted content from a web corpus. Although the classifier significantly decreased the presence of spam related words in the corpus, it was not able to recognise short non-text documents. That remains to be addressed in the future.

Acknowledgments

This work was partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2015071. This publication was written with the support of the Specific University Research provided by the Ministry of Education, Youth and Sports of the Czech Republic.

[5]Corpora in the study were lemmatised by TreeTagger.

[6]The tokenisation of the BNC had to be changed to the same way the web corpus was tokenised in order to make the counts of tokens in both corpora comparable.

[7]The comparison with the BNC also revealed there are words related to the modern technology (e.g. *website*, *online*, *email*) and American English spelled words (*center*, *organization*) in the 2015 web corpus.

References

[Baisa and Suchomel2014] Vít Baisa and Vít Suchomel. 2014. Skell: Web interface for english language learning. In *Eighth Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 63–70, Brno. Tribun EU.

[Baroni and Bernardini2004] Marco Baroni and Silvia Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *LREC*.

[Gyöngyi and Garcia-Molina2005] Zoltan Gyöngyi and Hector Garcia-Molina. 2005. Web spam taxonomy. In *First international workshop on adversarial information retrieval on the web (AIRWeb 2005)*.

[Joulin et al.2016] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

[Kilgarriff and Suchomel2013] Adam Kilgarriff and Vít Suchomel. 2013. Web spam. In Paul Rayson Stefan Evert, Egon Stemle, editor, *Proceedings of the 8th Web as Corpus Workshop (WAC-8) @Corpus Linguistics 2013*, pages 46–52.

[Marek et al.2007] Michal Marek, Pavel Pecina, and Miroslav Spousta. 2007. Web page cleaning with conditional random fields. In *Building and Exploring Web Corpora: Proceedings of the Fifth Web as Corpus Workshop, Incorporationg CleanEval (WAC3), Belgium*, pages 155–162.

[Ntoulas et al.2006] Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. 2006. Detecting spam web pages through content analysis. In *Proceedings of the 15th international conference on World Wide Web*, pages 83–92. ACM.

[Pomikálek2011] Jan Pomikálek. 2011. *Removing boilerplate and duplicate content from web corpora*. Ph.D. thesis, Masaryk University.

[Schäfer and Bildhauer2013] Roland Schäfer and Felix Bildhauer. 2013. *Web Corpus Construction*, volume 6. Morgan & Claypool Publishers.

[Spoustová and Spousta2012] Johanka Spoustová and Miroslav Spousta. 2012. A high-quality web corpus of Czech. In *LREC*, pages 311–315.

[Versley and Panchenko2012] Yannick Versley and Yana Panchenko. 2012. Not just bigger: Towards better-quality web corpora. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pages 44–52.

Table 1: Comparison of the corpus before and after spam removal using the classifier. Corpus sizes and relative frequencies (number of occurrences per million words) of selected words are shown. Reducing the corpus to 55 % of the former token count, phrases strongly indicating spam documents such as "cialis 20 mg", "payday loan" or "essay writing" were almost removed while innocent phrases from the same domains such as "oral administration", "interest rate" or "pass the exam" were reduced proportionally to the whole corpus.

	Original corpus	**Cleaned corpus**	**Kept in cleaned**
Document count	58,438,034	37,810,139	64.7 %
Token count	33,144,241,513	18,371,812,861	55.4 %
"viagra"	229.71	3.42	0.8 %
"cialis 20 mg"	2.74	0.02	0.4 %
"aspirin"	5.63	1.52	14.8 %
"oral administration"	0.26	0.23	48.8 %
"loan"	166.32	48.34	16.1 %
"payday loan"	24.19	1.09	2.5 %
"cheap"	295.31	64.30	12.1 %
"interest rate"	14.73	9.80	36.7 %
"essay"	348.89	33.95	5.4 %
"essay writing"	7.72	0.32	2.3 %
"pass the exam"	0.34	0.36	59.4 %

Figure 2: Relative wordcount comparison of the original 2015 web corpus with British National Corpus, top 26 lemmas sorted by the keyword score. Score = $\frac{fpm_1+100}{fpm_2+100}$ where fpm_1 is the count of lemmas per million in the focus corpus (3rd column) and fpm_2 is the count of lemmas per million in the reference corpus (5th column).

	Original English Web 2015		*British National Corpus*		
Lowercase lemma	**frequency**	**frequency/mill**	**frequency**	**frequency/mill**	**Score**
download	32,877,718	992.0	35	0.3	10.9
pdf	30,658,156	925.0	37	0.3	10.2
online	23,683,595	714.6	596	5.3	7.7
program	20,333,705	613.5	5,814	51.8	4.7
website	9,586,380	289.2	0	0.0	3.9
center	9,903,586	298.8	573	5.1	3.8
essay	11,563,807	348.9	2,317	20.6	3.7
viagra	7,620,095	229.9	0	0.0	3.3
url	7,168,836	216.3	0	0.0	3.2
ebook	6,969,380	210.3	0	0.0	3.1
web	7,206,520	217.4	729	6.5	3.0
internet	6,248,400	188.5	97	0.9	2.9
student	24,584,996	741.8	22,133	197.1	2.8
cialis	5,816,475	175.5	0	0.0	2.8
blog	5,110,812	154.2	0	0.0	2.5
email	5,074,946	153.1	43	0.4	2.5
cheap	9,787,744	295.3	6,649	59.2	2.5
epub	4,761,306	143.7	0	0.0	2.4
video	10,278,042	310.1	7,672	68.3	2.4
free	20,406,767	615.7	21,963	195.6	2.4
u.s.	4,976,297	150.1	458	4.1	2.4
post	13,400,787	404.3	12,576	112.0	2.4
outlet	5,501,465	166.0	1,375	12.2	2.4
color	4,553,463	137.4	143	1.3	2.3
click	5,326,832	160.7	1,273	11.3	2.3
your	95,303,049	2875.4	134,413	1197.0	2.3

Figure 3: Relative wordcount comparison of the cleaned web corpus with British National Corpus

	Cleaned English Web 2015		*British National Corpus*		
Lowercase lemma	**frequency**	**frequency/mill**	**frequency**	**frequency/mill**	**Score**
program	14,384,115	782.9	5,814	51.8	5.8
center	7,509,618	408.8	573	5.1	4.8
website	4,792,518	260.9	0	0.0	3.6
student	16,973,541	923.9	22,133	197.1	3.4
online	4,753,580	258.7	596	5.3	3.4
u.s.	4,225,425	230.0	458	4.1	3.2
project	14,949,773	813.7	21,742	193.6	3.1
university	12,182,707	663.1	18,899	168.3	2.8
community	15,164,485	825.4	26,564	236.6	2.7
global	4,585,347	249.6	3,529	31.4	2.7
web	3,322,320	180.8	729	6.5	2.6
download	3,011,631	163.9	35	0.3	2.6
email	2,901,189	157.9	43	0.4	2.6
dr.	3,290,385	179.1	1,215	10.8	2.5
internet	2,753,028	149.9	97	0.9	2.5
our	39,914,081	2172.6	93,457	832.3	2.4
click	3,144,338	171.2	1,273	11.3	2.4
focus	6,345,601	345.4	9,538	84.9	2.4
technology	7,397,599	402.7	12,865	114.6	2.3
organization	5,944,514	323.6	9,240	82.3	2.3
research	12,854,262	699.7	27,567	245.5	2.3
update	3,452,461	187.9	2,814	25.1	2.3
datum	7,682,640	418.2	14,212	126.6	2.3
network	5,810,016	316.2	9,291	82.7	2.3
video	5,202,487	283.2	7,672	68.3	2.3
photo	3,054,229	166.2	2,036	18.1	2.3

Figure 4: Relative wordcount comparison of the original web corpus with the cleaned version

	Original English Web 2015		*Cleaned English Web 2015*		
Lowercase lemma	**frequency**	**frequency/mill**	**frequency**	**frequency/mill**	**Score**
pdf	30,658,156	925.0	1,851,347	100.8	5.1
download	32,877,718	992.0	3,011,631	163.9	4.1
essay	11,563,807	348.9	623,760	34.0	3.4
viagra	7,620,095	229.9	62,899	3.4	3.2
ebook	6,969,380	210.3	265,781	14.5	2.7
cialis	5,816,475	175.5	45,613	2.5	2.7
url	7,168,836	216.3	509,596	27.7	2.5
buy	17,364,124	523.9	2,867,958	156.1	2.4
cheap	9,787,744	295.3	1,180,506	64.3	2.4
online	23,683,595	714.6	4,753,580	258.7	2.3
epub	4,761,306	143.7	203,405	11.1	2.2
prescription	4,646,919	140.2	280,013	15.2	2.1
outlet	5,501,465	166.0	651,024	35.4	2.0
book	29,921,305	902.8	7,889,796	429.5	1.9
generic	3,594,096	108.4	257,090	14.0	1.8
ugg	3,022,464	91.2	93,591	5.1	1.8
loan	5,512,504	166.3	888,181	48.3	1.8
jersey	4,873,552	147.0	836,729	45.5	1.7
insurance	7,150,681	215.7	1,588,816	86.5	1.7
pharmacy	2,941,876	88.8	290,211	15.8	1.6
sex	6,452,251	194.7	1,502,817	81.8	1.6
de	10,572,331	319.0	2,986,557	162.6	1.6
mg	2,776,320	83.8	298,031	16.2	1.6
you	195,234,032	5890.4	68,409,350	3723.6	1.6
binary	4,226,875	127.5	839,475	45.7	1.6
levitra	1,873,011	56.5	34,646	1.9	1.5